Protocol I4V-MC-JAIY(a)
A Multicenter, Randomized, Double-Blind, Placebo-Controlled, Phase 3 Study to Evaluate the Efficacy and Safety of Baricitinib in Combination with Topical Corticosteroids in Adult Patients with Moderate to Severe Atopic Dermatitis
BREEZE-AD7

EudraCT number 2018-001726-26

Baricitinib (LY3009104)

Eli Lilly and Company
Indianapolis, Indiana USA 46285

Protocol Electronically Signed and Approved by Lilly on 05 July 2018.
Amendment (a) Electronically Signed and Approved by Lilly on Date Provided Below.

Approval Date: 18-Dec-2018 GMT

Table of Contents

Section **Page**

List of Tables

Table **Page**

List of Figures

Figure **Page**

List of Appendices

1. Synopsis

Title of Study:
A Multicenter, Randomized, Double-Blind, Placebo-Controlled, Phase 3 Study to Evaluate the Efficacy and Safety of Baricitinib in Combination with Topical Corticosteroids in Adult Patients with Moderate to Severe Atopic Dermatitis - BREEZE-AD7

Rationale:
Atopic dermatitis (AD) is a pruritic, chronic or chronically relapsing, highly symptomatic inflammatory skin disease characterized by excessive T cell activation leading to significant skin infiltration by T cells and dendritic cells (Bieber 2010). Presentation is varied, but includes skin manifestations and pruritus, with associated sleep disturbances and subsequent skin infections. The course of disease includes relapses of varying duration and severity.

Baricitinib is an orally available, selective Janus kinase (JAK) inhibitor with potency and selectivity for JAK1 and JAK2 and less potency for JAK3 or tyrosine kinase 2 (TYK2) (Fridman et al. 2010). The pathogenesis of AD is thought to be modulated through thymic stromal lymphopoietin (TSLP), interleukin (IL)-13, IL-4, IL-5, IL-22, and IL-31, many of which activate receptors with downstream signaling through intracellular JAK1/JAK2/TYK2 (Nomura and Kabashima 2015). This activity profile suggests that baricitinib would inhibit cytokines involved in AD pathogenesis.

Clinical studies have established that baricitinib is effective in autoimmune/autoinflammatory diseases involving the joints, kidneys, and skin. Baricitinib was effective at reducing swollen and tender joints in patients with rheumatoid arthritis (Genovese et al. 2016; Dougados et al. 2017; Fleischmann et al. 2017; Taylor et al. 2017); was effective at reducing disease severity in patients with moderate to severe plaque psoriasis (Papp et al. 2016); was effective at reducing the urinary albumin-to-creatinine ratio in patients with diabetic kidney disease (Tuttle et al. 2015); and in a recently completed Phase 2 study (I4V-MC-JAHG) was effective at reducing disease severity in patients with moderate to severe AD (Guttman-Yassky et al. 2018). The mechanism of action, combined with demonstration of clinical benefits in inflammatory diseases involving joints, kidneys, and skin, provides the rationale for evaluating baricitinib in moderate–to-severe AD.

Objectives/Endpoints:

Objectives	Endpoints
Primary	
To test the hypothesis that baricitinib 4-mg QD + TCS or baricitinib 2-mg QD + TCS is superior to placebo + TCS in the treatment of patients with moderate to severe AD.	• Proportion of patients achieving IGA of 0 or 1 with a ≥2-point improvement at Week 16.
Key Secondary *These are prespecified objectives that will be adjusted for multiplicity*	
To compare the efficacy of baricitinib 2-mg QD + TCS or baricitinib 4-mg QD + TCS to placebo + TCS in AD during the 16-week double-blind placebo-controlled treatment period as measured by improvement in signs and symptoms of AD.	• Proportion of patients achieving EASI75 at 16 weeks • Proportion of patients achieving EASI90 at 16 weeks • Percent change from baseline in EASI score at 16 weeks • Proportion of patients achieving SCORAD75 at 16 weeks.

Objectives	Endpoints
To compare the efficacy of baricitinib 2-mg QD + TCS or baricitinib 4-mg QD + TCS to placebo + TCS in AD during the 16-week double-blind placebo-controlled treatment period as assessed by patient-reported outcome measures.	• Proportions of patients achieving a 4-point improvement in Itch NRS at 2 days, 1 week, 2 weeks, 4 weeks, and 16 weeks • Mean change from baseline in the score of Item 2 of the ADSS at 1 week and 16 weeks • Mean change from baseline in Skin Pain NRS at 16 weeks.
Other Secondary Objectives *These are prespecified objectives that will not be adjusted for multiplicity.*	
To compare the efficacy of baricitinib 2-mg QD + TCS or baricitinib 4-mg QD + TCS to placebo + TCS in AD during the 16-week double-blind placebo-controlled period as measured by improvement in signs and symptoms of AD.	• Proportion of patients achieving IGA of 0 or 1 with a ≥2-point improvement at Week 4 • Proportion of patients achieving EASI50 at 16 weeks • Proportion of patients achieving IGA of 0 at 16 weeks • Mean change from baseline in SCORAD at 16 weeks • Proportion of patients achieving SCORAD90 at 16 weeks • Mean change from baseline in body surface area affected at 16 weeks • Proportion of patients developing skin infections requiring antibiotic treatment by Week 16 • Mean gram quantity of background TCS used over 16 weeks (tube weights)
To compare the efficacy of baricitinib 2-mg QD + TCS or baricitinib 4-mg QD + TCS to placebo + TCS in AD during the 16-week, double-blind, placebo-controlled treatment period as assessed by patient-reported outcome/QoL measures.	• Percent change from baseline in Itch NRS at 2 days, 1 week, 4 weeks, and 16 weeks • Mean change from baseline in Itch NRS at 2 days, 1 week, 4 weeks, and 16 weeks • Mean change from baseline in the total score of the POEM at 16 weeks • Mean change in PGI-S-AD scores at 16 weeks • Mean change from baseline in the HADS at 16 weeks • Mean change in DLQI scores at 16 weeks • Mean change in WPAI scores at 16 weeks • Mean change in EQ-5D-5L scores at 16 weeks. • Mean number of days without use of background TCS over 16 weeks

Abbreviations: AD = atopic dermatitis; ADSS = Atopic Dermatitis Sleep Scale; DLQI = Dermatology Life Quality Index; EASI = Eczema Area and Severity Index; EQ-5D-5L = the European Quality of Life–5 Dimensions–5 Levels; HADS = Hospital Anxiety Depression Scale; IGA = Investigator's Global Assessment; NRS = numeric rating scale; QD = once daily; PGI-S-AD = Patient Global Impression of Severity–Atopic Dermatitis; POEM = Patient-Oriented Eczema Measure; QoL = quality of life; SCORAD = SCORing Atopic Dermatitis; TCS = topical corticosteroids; WPAI = Work Productivity and Activity Impairment.

Summary of Study Design:

Study I4V-MC-JAIY (JAIY) is a Phase 3, multicenter, randomized, double-blind, placebo-controlled, parallel-group, outpatient study evaluating the efficacy and safety of baricitinib 4-mg once daily (QD) plus topical corticosteroids (TCS) and 2-mg QD plus TCS, as compared to placebo plus TCS in adult patients with moderate to severe AD. The study population will include patients aged 18 years or older who have moderate to severe AD and a history of inadequate response to existing topical therapies.

The study duration will be up to 25 weeks over 3 study periods:

- Period 1: Screening Period lasting from 8 to 35 days prior to Week 0 (baseline, Visit 2).
- Period 2: Double-Blinded Treatment Period, lasting from Week 0 (baseline, Visit 2) through Week 16 (Visit 8) inclusive:
    - At completion of the double-blind treatment period, eligible patients will be provided the option to participate in the long-term extension study I4V-MC-JAHN (JAHN). Those not eligible or who chose not to participate will proceed to the post-treatment follow-up period.
- Period 3: Post-Treatment Follow-Up Period, spanning the period from the last treatment visit at Week 16 (Visit 8) or Early Termination Visit (ETV) to approximately 4 weeks following the last dose of investigational product.

Treatment Arms and Duration

Patients will be randomized at Week 0 to 1 of 3 treatment groups: placebo once daily (QD), baricitinib 2-mg QD, or baricitinib 4-mg QD. Use of TCS as a background therapy is allowed during the study. The study duration will be up to 25 weeks (Screening Period: up to 5 weeks prior to randomization; Double-Blinded Treatment Period: 16 weeks; Follow-up Period: approximately 4 weeks after the last dose of investigational product).

Number of Patients

This study will include approximately 300 patients with AD who will be randomized in a 1:1:1 ratio to receive placebo QD, baricitinib 2-mg QD or baricitinib 4-mg QD (100 patients in each treatment group).

Statistical Analysis

Unless otherwise specified, the efficacy and health outcome analyses will be conducted on the intent-to-treat population and safety analyses will be conducted on those patients who receive at least 1 dose of investigational product.

Treatment comparisons of discrete efficacy variables will be made using a logistic regression analysis with treatment, baseline disease severity, and region in the model. The proportions and 95% confidence interval (CI) will be reported. If a patient needs to use rescue medication, the data after rescue onward will be considered missing and missing data will be imputed using the nonresponder imputation (NRI) method. All patients who discontinue the study or study treatment at any time for any reason will be defined as nonresponders for the NRI analysis for categorical variables after discontinuation onward. Additional analyses will be done using all observed data whether rescue medication was used or not.

Treatment comparisons of continuous efficacy and health outcome variables will be made using mixed-effects model of repeated measures (MMRM) with treatment, region, baseline severity, visit, and treatment-by-visit interaction as fixed categorical effects and baseline score and baseline score-by-visit interaction as fixed continuous effects. An unstructured covariance matrix will be used to model the within-patient variance–covariance errors. Type III sums of squares for the least squares means (LSMs) will be used for the statistical comparison and contrasts will be set up within the model to compare treatment groups at specific time points of interest.

Fisher exact test will be used for all adverse events (AEs), baseline, discontinuation, and other categorical safety data. Continuous vital signs and laboratory values will be analyzed by an analysis of covariance (ANCOVA) with treatment and baseline values in the model.

2. Schedule of Activities

Table JAIY.1. I4V-MC-JAIY Schedule of Activities

	Screening	Double-Blind, Placebo-Controlled							Post-treatment Follow-up
	Period 1	Period 2							Period 3
Visit Number	1	2	3	4	5	6	7	8	801[a]
Weeks from Randomization		0	1	2	4	8	12	16 or ET	
Days from Randomization		0	7	14	28	56	84	112	
Visit Tolerance Interval (days)	-8 to -35		±2	±2	±2	±4	±4	±4	28 ± 4 after last dose
Procedure									
Inclusion and exclusion review	X	X							
Informed consent	X								
Clinical Assessments									
Demographics	X								
Medical history	X								
Substance Use (alcohol, tobacco use)	X								
Previous and current AD treatments	X								
Weight	X	X			X	X	X	X	X
Height		X							
Vital signs (BP and pulse rate)	X	X	X	X	X	X	X	X	X
Physical examination	X								
Symptom-directed physical examination[b]		X	X	X	X	X	X	X	X
12-lead ECG (single)	X								
Chest x-ray[c] (posterior–anterior view)	X								
TB test[d]	X								
Read PPD if applicable (48-72 hours post PPD)	X[e]								
Pre-existing Conditions	X								
Adverse Events		X	X	X	X	X	X	X	X
Concomitant Medication	X	X	X	X	X	X	X	X	X
ePRO (patient diary) dispensed	X	X	X	X	X	X	X	X[g]	
ePRO (patient diary) returned[f]		X	X	X	X	X	X	X	X[g]

	Screening	Double-Blind, Placebo-Controlled							Post-treatment Follow-up
	Period 1	Period 2							Period 3
Visit Number	1	2	3	4	5	6	7	8	801[a]
Weeks from Randomization		0	1	2	4	8	12	16 or ET	
Days from Randomization		0	7	14	28	56	84	112	
Visit Tolerance Interval (days)	-8 to -35		±2	±2	±2	±4	±4	±4	28 ± 4 after last dose
Randomization		X							
IWRS	X	X	X	X	X	X	X	X	X
IP dispensed		X	X	X	X	X	X		
IP returned and compliance assessed			X	X	X	X	X	X	
Dispense background TCS[h]		X	X	X	X	X	X	X	
Weigh (tube with cap) and record returned background TCS[h]			X	X	X	X	X	X	X
Scales									
IGA	X	X	X	X	X	X	X	X	X
EASI	X	X	X	X	X	X	X	X	X
SCORAD	X	X	X	X	X	X	X	X	X
Health Outcome Measures and Other Questionnaires[i]									
POEM	X	X	X	X	X	X	X	X	X
DLQI	X	X	X	X	X	X		X	X
HADS	X	X		X	X	X		X	
EQ-5D-5L		X	X	X	X	X	X	X	X
WPAI-AD		X	X	X	X	X	X	X	X
Itch NRS	X	X	X	X	X	X	X	X	X
Skin Pain NRS	X	X	X	X	X	X	X	X	X
ADSS	X	X	X	X	X	X	X	X	X
PGI-S-AD	X	X	X	X	X	X	X	X	X
PIQ – General[j]	X	X	X	X	X	X	X	X	X
PIQ – Activity and Clothing[j]	X	X	X	X	X	X	X	X	X
PIQ – Mood and Sleep[j]	X	X	X	X	X	X	X	X	X
PIQ – Scratching Behavior[j]	X	X	X	X	X	X	X	X	X

	Screening	Double-Blind, Placebo-Controlled							Post-treatment Follow-up
	Period 1	Period 2							Period 3
Visit Number	1	2	3	4	5	6	7	8	801[a]
Weeks from Randomization		0	1	2	4	8	12	16 or ET	
Days from Randomization		0	7	14	28	56	84	112	
Visit Tolerance Interval (days)	-8 to -35		±2	±2	±2	±4	±4	±4	28 ± 4 after last dose
PROMIS – Sleep-Related Impairment[j]	X	X	X	X	X	X	X	X	X
Neuro-QoL – Cognitive Function[j]	X	X						X	
Patient Benefit Index[j]	X	X						X	
C-SSRS[k]/Self-Harm Supplement Form	X	X	X	X	X	X	X	X	X
Self-Harm Follow-up Form[l]	X	X	X	X	X	X	X	X	X
Laboratory Assessment									
Clinical Chemistry[m]	X	X			X	X	X	X	X
Hematology	X	X			X	X	X	X	X
Lipids (fasting)[n]		X					X	X[o]	X
Serum Pregnancy[p]	X								
FSH[q]	X								
TSH	X								
HIV	X								
HCV antibody[r]	X								
HBV testing	X								
HBV DNA[s]	X							X	
Urinalysis	X	X			X	X	X	X	X
Urine Pregnancy[p]		X		X	X	X	X	X	X
Pharmacogenetics: blood		X							
Serum immunoglobulins		X			X			X	
Exploratory storage samples (serum and plasma)		X			X			X	
RNA and biomarkers: blood		X			X			X	

Abbreviations: AD = atopic dermatitis; ADSS = Atopic Dermatitis Sleep Scale; BP = blood pressure; C-SSRS = Columbia Suicide Severity Rating Scale 11 categories suicidal ideation/suicidal behavior; DLQI = Dermatology Life Quality Index; DNA = deoxyribonucleic acid; EASI = Eczema Area and Severity Index; ECG = electrocardiogram; eGFR = estimated glomerular filtration rate; EQ-5D-5L = the European Quality of Life–5 Dimensions–5 Levels; ET = early termination; ePRO = electronic patient-reported outcomes (device); ETV = early termination visit; FSH = follicle-stimulating hormone; HADS = Hospital Anxiety Depression Scale; HBcAb = hepatitis B core antibody; HBsAb = hepatitis B surface antibody; HBV = hepatitis B virus; HCV = hepatitis C virus; HIV = human immunodeficiency virus; IGA = Investigator's Global Assessment; IP = investigational product; IWRS = interactive web-response system; Neuro-QoL = Quality of Life in Neurological Disorders; NRS = numeric rating scale ; PBI = Patient Benefit Index; PGI-S-AD = Patient Global Impression of Severity–Atopic Dermatitis; PIQ = Pain Impact Questionnaire; POEM = Patient-Oriented Eczema Measure; PPD = purified protein derivative; PROMIS = Patient-Reported Outcomes Measurement Information System; RNA = ribonucleic acid; SCORAD = SCORing Atopic Dermatitis; TB = tuberculosis; TCS = topical corticosteroids; TSH =thyroid-stimulating hormone; WPAI-AD = Work Productivity and Activity Impairment–Atopic Dermatitis.

[a] Patients who have discontinued IP, but remain in the study for more than 28 days without IP can combine their Visit 8/ET visit with their Visit 801 (follow-up visit).

[b] The symptom-directed physical examination may be repeated at the investigator's discretion any time a patient presents with physical complaints.

[c] A posterior–anterior chest x-ray will be performed at screening unless one has been performed within the past 6 months and the x-ray and reports are available.

[d] TB test(s) including PPD, QuantiFERON®-TB Gold, and T SPOT®. See Exclusion Criterion [39] for description of TB testing. In countries where the QuantiFERON-TB Gold test or T-SPOT is available, either test may be used instead of the PPD TB test. The QuantiFERON-TB Gold test may be performed centrally (recommended/preferred) or locally; the T-SPOT must be performed locally. (Note: Exception: Patients with a history of active or latent TB who have documented evidence of appropriate treatment, have no history of re-exposure since their treatment was completed, and have a screening chest x-ray with no evidence of active TB may be enrolled if other entry criteria are met. Such patients would not be required to undergo the protocol-specific TB testing but must have a chest x-ray at screening.)

[e] If PPD testing was chosen to test for TB, then the patient must return and have the PPD test read 48 to 72 hours after Visit 1 (post-PPD).

[f] ePRO devices will need to be collected from screen fail patients.

[g] For patients not entering Study JAHN, their JAIY patient diary will continue to be dispensed at the final visit and returned at Visit 801.

[h] Only as required based on clinical symptoms.

[i] The following measures (POEM, DLQI, EQ-5D-5L, WPAI-AD, PIQ, PROMIS, Neuro-QoL, and PBI) should be completed prior to any clinical assessments being performed on days when study visits occur.

[j] These will be conducted in the countries where translations are available.

[k] Suicidal ideation and behavior subscales excerpt—Adapted for the assessment of 11 preferred ideation and behavior categories.

[l] The Self-Harm Follow-up Form is only required if triggered by the Self-Harm Supplement Form.

[m] Clinical chemistry will include the following value calculated by the central laboratory from serum creatinine: estimated glomerular filtration rate (eGFR, calculated using the Chronic Kidney Disease Epidemiology Collaboration [CKD-EPI] Creatinine 2009 equation).

[n] Fasting lipid profile: Patients should not eat or drink anything except water for 12 hours prior to sample collection. If a patient attends these visits in a nonfasting state, this will not be considered a protocol violation.

[o] Only required for patients completing an ETV prior to completion of Visit 7 (Week 12).

[p] For all women of child-bearing potential, a serum pregnancy test (central laboratory) will be performed at Visit 1. Urine pregnancy tests (local laboratory) will be performed at Visit 2 and at all subsequent study visits after Visit 3.

[q] For female patients ≥40 and <60 years of age who have had a cessation of menses for at least 12 months, an FSH test will be performed to confirm nonchildbearing potential (FSH ≥40 mIU/mL).

[r] For patients who are positive for HCV antibody, a follow-up test for HCV RNA will be performed automatically. Patients who are positive for HCV antibody and negative for HCV RNA may be enrolled.

[s] Patients who are positive for HBcAb and negative for HBV DNA may be enrolled. Any enrolled patient who is HBcAb positive, regardless of HBsAb status or level, must undergo HBV DNA testing per the schedule (Section 9.4.8).

3. Introduction

3.1. Background

Atopic dermatitis (AD), also known as atopic eczema, is a common, chronic, relapsing, highly symptomatic inflammatory skin disease (Bieber 2010). Patients with AD may present with skin lesions that can be acute with oozing, crusted, eroded vesicles or papules on erythematous plaques. Patients may also present with lesions that have a subacute appearance, with thick and excoriated plaques, or chronic appearance, with lichenified, slightly pigmented, excoriated plaques (Bieber 2010). Atopic dermatitis causes pruritus attacks throughout the day, which is the primary source of morbidity in this disorder (Simpson 2012). Pruritus often leads to an "itch–scratch" cycle, further compromising the epidermal barrier and resulting in dry skin, microbial colonization, and secondary infections (Krakowski et al. 2008); 36% of patients have reported that they often or always scratch until their skin bleeds (Langenbruch et al. 2014). Pruritus from AD can worsen during night time, resulting in sleep disturbances; approximately 27% of adult patients with AD experiencing sleep disturbance as a result of itching (Langenbruch et al. 2014). In adult patients with moderate to severe AD, sleep quality and latency were significantly associated with poor quality of life (QoL) (Yano et al. 2013).

In clinical practice, AD is classified as mild, moderate, or severe based on a variety of clinical features, including severity of skin lesions and pruritus, and extent of disease (body surface area [BSA] involved).

Until recently, there were no Food and Drug Administration (FDA)-approved systemic treatments for patients with moderate to severe AD, with the exception of systemic corticosteroids, and in the European Union, only cyclosporine had been approved for the treatment of patients with severe AD (Bieber and Straeter 2015). In 2017, Dupixent (dupilumab) injection, an IgG4 monoclonal antibody that inhibits interleukin (IL)-4 and IL-13, was approved by the FDA and the European Medicines Agency (EMA) for this patient population. A recently completed Phase 2 study (I4V-MC-JAHG [JAHG]) evaluated the safety and efficacy of baricitinib (a Janus kinase [JAK] inhibitor) in AD and results showed significant improvement in disease severity compared to placebo and no new safety concerns were identified (Guttman-Yassky et al. 2018).

In addition to AD, baricitinib has also been studied in Phase 3 in patients with rheumatoid arthritis (RA) and in Phase 2 in patients with diabetic nephropathy, moderate to severe psoriasis, and systemic lupus erythematosus.

Through 13 February 2018, baricitinib has been studied in approximately 548 healthy volunteers and 4673 patients have received baricitinib in clinical studies. Of these, more than 2700 patients have been treated with baricitinib for more than a year and more than 2100 patients have been treated with baricitinib for more than 2 years at doses of 2-mg once daily (QD) and/or 4-mg QD across the RA clinical program. Baricitinib has been administered as single doses ranging from 1- to 40-mg and as repeat oral doses ranging from 2- to 20-mg to healthy subjects. Baricitinib has also been administered to patients with RA at doses up to 15-mg daily for 4 weeks, 10-mg

daily for 24 weeks, 8-mg daily for 76 weeks, and lower doses up to 4-mg daily for up to approximately 5 years.

3.2. Study Rationale

The underlying cause of AD is not completely understood. Loss of function mutations in the gene for *filaggrin* (filament aggregating protein), a key protein in terminal differentiation of the epidermis contributing to barrier function, has been identified as the strongest genetic risk factor for AD in European populations (Palmer et al. 2006). At a cellular level, AD is characterized by excessive T cell activation caused by genetic and environmental factors, leading to significant skin infiltration by T cells and dendritic cells. The cytokine thymic stromal lymphopoietin (TSLP) is thought to act as a master switch that triggers the initiation and maintenance of AD (Moniaga et al. 2013; Ziegler et al. 2013). Overexpression of TSLP in keratinocytes, the most prevalent cell type in the skin, triggers robust itch-evoked scratching and the development of an AD-like skin phenotype in mice (Li et al. 2005). In addition to directly inducing itch by activating sensory neurons in the skin, TSLP also enhances maturation and differentiation of dendritic cells and naive CD4+ T cells and induces production of Th2-related cytokines involved in AD pathogenesis (Wilson et al. 2013; Divekar and Kita 2015). Thymic stromal lymphopoietin and other key cytokines involved in AD pathogenesis, such as IL-13, IL-5, IL-22, and IL-31, signal through receptors associated with intracellular JAK1/JAK2/tyrosine kinase 2 (TYK2) signaling (Ziegler et al. 2013; Nomura and Kabashima 2015).

Janus kinases are a family of tyrosine kinases that mediate cytokine receptor signaling through phosphorylation and activation of signal transducers and activators of transcription (STAT) proteins. There are 4 known JAK family members: JAK1, JAK2, JAK3, and TYK2 (Clark et al. 2014). The relative affinity of JAK inhibitors for different members of the JAK kinase family allows for differentiation of JAK inhibitors in relation to their enzymatic inhibitory profile. In vitro assays indicate that baricitinib is a selective inhibitor of JAKs with potency and selectivity for JAK1/2 and less potency for JAK3 or TYK2 (Fridman et al. 2010). The balanced JAK1/JAK2 inhibitory profile of baricitinib suggests that baricitinib will have the greatest modulatory effect in cytokines signaling through a JAK1/JAK2 heterodimer intracellularly (or a JAK1/JAK2/TYK2), such as IL-6, TSLP, IL-13, or IL-31 (Vaddi and Luchi 2012).

The recently completed Phase 2 study of baricitinib in AD, JAHG, met its primary objective of proportion of patients achieving a 50% improvement from baseline in Eczema Area and Severity Index (EASI) scores compared to placebo (Guttman-Yassky et al. 2018). Baricitinib also showed statistically significant improvements for other disease severity analyses as well as multiple different patient-reported outcome (PRO) scales compared to placebo, further validating the hypothesis that JAK1/JAK2 signaling plays a key role in AD pathogenesis.

To comprehensively evaluate the efficacy of baricitinib in patients with AD, both as a monotherapy and in combination with background TCS, several Phase 3 studies have been initiated. Two multiregional Phase 3 studies (I4V-MC-JAHL [BREEZE-AD1] and I4V-MC-JAHM [BREEZE-AD2]) will evaluate the safety and efficacy of baricitinib monotherapy as compared to placebo, in adult patients with moderate to severe AD. A

long-term extension study I4V-MC-JAHN (JAHN; discussed in Sections 5.1.2 and 7.8.1) follows studies JAHL, JAHM, and this study (I4V-MC-JAIY [JAIY]), and includes a randomized treatment withdrawal and downtitration substudy. Study I4V-MC-JAIN (BREEZE- AD4) is a multiregional Phase 3 study evaluating the safety and efficacy of baricitinib in combination with TCS in patients with moderate to severe AD who have experienced failure of cyclosporine, or are intolerant to, or have a contraindication to, cyclosporine.

This study, JAIY (BREEZE-AD7), is designed in the same patient population as studies JAHL and JAHM. However, JAIY will assess baricitinib in combination with TCS, and will provide additional information on timing and impact of baricitinib on patient-reported itch, a hallmark of AD.

3.3. Benefit/Risk Assessment

CCI

[redacted]

CCI

More information about the known and expected benefits, risks, serious adverse events (SAEs), and reasonably anticipated adverse events (AEs) of baricitinib are to be found in the investigator's brochure (IB).

4. Objectives and Endpoints

Table JAIY.2 shows the objectives and endpoints of the study.

Table JAIY.2. Objectives and Endpoints

Objectives	Endpoints
Primary Objective	
To test the hypothesis that baricitinib 4-mg QD + TCS or baricitinib 2-mg QD + TCS is superior to placebo + TCS in the treatment of patients with moderate to severe AD.	• Proportion of patients achieving IGA of 0 or 1 with a ≥2-point improvement at Week 16.
Key Secondary Objectives *These are prespecified objectives that will be adjusted for multiplicity*	
To compare the efficacy of baricitinib 2-mg QD + TCS or baricitinib 4-mg QD + TCS to placebo + TCS in AD during the 16-week double-blind placebo-controlled treatment period as measured by improvement in signs and symptoms of AD.	• Proportion of patients achieving EASI75 at 16 weeks • Proportion of patients achieving EASI90 at 16 weeks • Percent change from baseline in EASI score at 16 weeks • Proportion of patients achieving SCORAD75 at 16 weeks.
To compare the efficacy of baricitinib 2-mg QD + TCS or baricitinib 4-mg QD + TCS to placebo + TCS in AD during the 16-week double-blind placebo-controlled treatment period as assessed by patient-reported outcome measures.	• Proportions of patients achieving a 4-point improvement in Itch NRS at 2 days, 1 week, 2 weeks, 4 weeks, and 16 weeks • Mean change from baseline in the score of Item 2 of the ADSS at 1 week and 16 weeks. • Mean change from baseline in Skin Pain NRS at 16 weeks.
Other Secondary Objectives *These are prespecified objectives that will not be adjusted for multiplicity.*	
To compare the efficacy of baricitinib 2-mg QD + TCS or baricitinib 4-mg QD + TCS to placebo + TCS in AD during the 16-week double-blind placebo-controlled period as measured by improvement in signs and symptoms of AD.	• Proportion of patients achieving IGA of 0 or 1 with a ≥2-point improvement at Week 4 • Proportion of patients achieving EASI50 at 16 weeks • Proportion of patients achieving IGA of 0 at 16 weeks • Mean change from baseline in SCORAD at 16 weeks • Proportion of patients achieving SCORAD90 at 16 weeks • Mean change from baseline in body surface area affected at 16 weeks • Proportion of patients developing skin infections requiring antibiotic treatment by Week 16. • Mean gram quantity of background TCS used over 16 weeks (tube weights)

Objectives	Endpoints
To compare the efficacy of baricitinib 2-mg QD + TCS or baricitinib 4-mg QD + TCS to placebo + TCS in AD during the 16-week, double-blind, placebo-controlled treatment period as assessed by patient-reported outcome/QoL measures.	• Percent change from baseline in Itch NRS at 2 days, 1 week, 4 weeks, and 16 weeks • Mean change from baseline in Itch NRS at 2 days, 1 week, 4 weeks, and 16 weeks • Mean change from baseline in the total score of the POEM at 16 weeks • Mean change in PGI-S-AD scores at 16 weeks • Mean change from baseline in the HADS at 16 weeks • Mean change in DLQI scores at 16 weeks • Mean change in WPAI scores at 16 weeks • Mean change in EQ-5D-5L scores at 16 weeks • Mean number of days without use of background TCS over 16 weeks

Exploratory Objectives/Endpoints

- Frequency of patient-reported "no itch" (Itch NRS score = 0) days from daily diaries from Week 12 to Week 16
- Frequency of patient-reported "no pain" (Skin Pain NRS score = 0) days from daily diaries from Week 12 to Week 16
- Mean change from baseline in PIQ – General score
- Mean change from baseline in PIQ – Activity and Clothing score
- Mean change from baseline in PIQ – Mood and Sleep score
- Mean change from baseline in PIQ – Scratching Behavior score
- Mean change from baseline in PROMIS – Sleep-Related Impairment score
- Mean change from baseline in Neuro-QoL – Cognitive Function score
- Mean change from baseline in Patient Benefit Index score – global score plus the following subscales:
    - Reducing social impairments
    - Reducing psychological impairments
    - Reducing impairments due to therapy
    - Reducing physical impairments
    - Having confidence in healing.
- Mean change from baseline in the score of Item 1 of the ADSS at 1 week and 16 weeks
- Mean change from baseline in the score of Item 3 of the ADSS at 1 week and 16 weeks
- To evaluate changes from baseline in IgE levels during the study
- To evaluate changes from baseline in eosinophil levels during the study

Abbreviations: AD = atopic dermatitis; ADSS = Atopic Dermatitis Sleep Scale; DLQI = Dermatology Life Quality Index; EASI = Eczema Area and Severity Index; EQ-5D-5L = the European Quality of Life–5 Dimensions–5 Levels; HADS = Hospital Anxiety Depression Scale; IgE = immunoglobulin E; IGA = Investigator's Global Assessment; Neuro-QoL = Quality of Life in Neurological Disorders; NRS = numeric rating scale; PIQ = Pain Impact Questionnaire; PROMIS = Patient-Reported Outcomes Measurement Information System; QD = once daily; QoL = quality of life; PGI-S-AD = Patient Global Impression of Severity–Atopic Dermatitis; POEM = Patient-Oriented Eczema Measure; SCORAD = SCORing Atopic Dermatitis; TCS = topical corticosteroids; WPAI = Work Productivity and Activity Impairment.

5. Study Design

5.1. Overall Design

Study JAIY is a Phase 3, multicenter, randomized, double-blind, placebo-controlled, parallel-group, outpatient study evaluating the efficacy and safety of baricitinib 2-mg QD and 4-mg QD, in combination with TCS, as compared to placebo in combination with TCS, in adult patients with moderate to severe AD. The study is divided into 3 periods, a 5-week Screening period, a 16-week Double-Blinded Treatment period, and a 4-week Post-Treatment Follow-Up period. For those patients who complete the 16-week treatment period, there is an option to participate in the long-term extension Study JAHN.

Approximately 300 patients ≥18 years of age who have responded inadequately to topical therapy will be randomized in a 1:1:1 ratio to receive placebo QD, baricitinib 2-mg QD, or baricitinib 4-mg QD in combination with TCS (100 patients in each treatment group). Patients will be stratified at randomization according to disease severity (Investigator's Global Assessment [IGA] 3 vs. 4) and geographic region.

All procedures to be conducted during the study, including timing of all procedures, are indicated in the Schedule of Activities (Section 2). Section 9.4.4 describes collection of laboratory samples; Appendix 2, Appendix 4, and Appendix 5 list the specific laboratory tests that will be performed for this study. Laboratory samples listed in Appendix 4, Appendix 5, and Appendix 6 are collected when possible in the event of specific AEs. Study governance considerations are described in detail in Appendix 3. Section 10.3.7.1 outlines information regarding the data monitoring committee (DMC) and interim analyses.

5.1.1. Period 1: Screening

The duration of the Screening Period is between 8 and 35 days prior to Visit 2 (Week 0). At Visit 1, the patient will sign the informed consent form (ICF) prior to any study assessments, examinations, or procedures being performed (Appendix 3). All screening procedures will be performed according to the Schedule of Activities (Section 2). Patients who receive a purified protein derivative (PPD) skin test at Visit 1 will need to return within 48 to 72 hours later to read the skin test. Prior to randomization, treatments for AD will be washed out: 4 weeks for systemic treatments and 2 weeks for topical treatments (not including emollients). Patients will be required to use emollients daily during the 14 days preceding randomization and throughout the study. If patients have been using emollients daily at the time of screening, then those cumulative days can be utilized to meet inclusion criterion [8]. Additionally, collection of data through daily diaries will be required throughout the screening period. The baseline for the daily PRO assessments will be the average score of the 7 days prior to randomization; thus, the minimum screening window was set at 8 days.

All patients who have not previously received the herpes zoster vaccine by screening will be encouraged (per local guidelines) to do so prior to randomization. Refer to the exclusion criterion [28] in Section 6 for additional information regarding herpes zoster vaccinations. In addition, investigators should review the vaccination status of their patients and follow the local

guidelines for vaccination of those >18 years of age with nonlive vaccines intended to prevent infectious disease prior to entering patients into the study.

Patients who meet all of the inclusion and none of the exclusion criteria (Section 6) will continue to Visit 2.

5.1.2. Period 2: Double–Blind, Placebo-Controlled Treatment

At Visit 2 (Week 0, baseline), study eligibility for each patient will be reviewed, based on all inclusion and exclusion criteria (Section 6), and laboratory test results. Patients who meet all criteria will proceed to randomization and begin the 16-week double-blind, placebo-controlled treatment period.

At Visit 2, after laboratory samples are collected and all assessments are completed, patients will take the first dose of investigational product at the clinic.

Patient will be randomized at a 1:1:1 ratio into 1 of the 3 treatment groups (placebo QD, baricitinib 2-mg QD, or baricitinib 4-mg QD). Investigational product will be administered daily for 16 weeks (treatment period Visits 2 through 8; Section 7). All patients will be required to use emollients daily. Daily diaries will continue to be utilized throughout the treatment period. Download of this data will be required at study visits. TCS will be dispensed at V2 and used on affected areas as described in section 7.7.2. Topical calcineurin inhibitors (TCNIs) is also allowed, but TCNI use should be limited to problem areas (e.g. face and skin folds). The use of higher potency TCS and systemic therapies for the treatment of AD are not allowed, except as part of rescue therapy for patients not responding to treatment. Details of background topical therapy, as well as rescue therapy and rescue criteria are included in Section 7.7. Assessments of disease severity will be performed by the investigator at all study visits including unscheduled and early termination visits (ETVs).

The primary efficacy endpoint and final visit in the treatment period will be at Week 16 (Visit 8). Patients who complete through the Week 16 study visit may be eligible for inclusion in the long-term extension study JAHN (up to 2 additional years of treatment).

If a patient discontinues investigational product for any reason, the patient should remain in the study through Week 16 (Visit 8). If the patient refuses and wishes to withdraw consent, an ETV should be performed as soon as logistically possible.

5.1.3. Period 3: Post-Treatment Follow-up

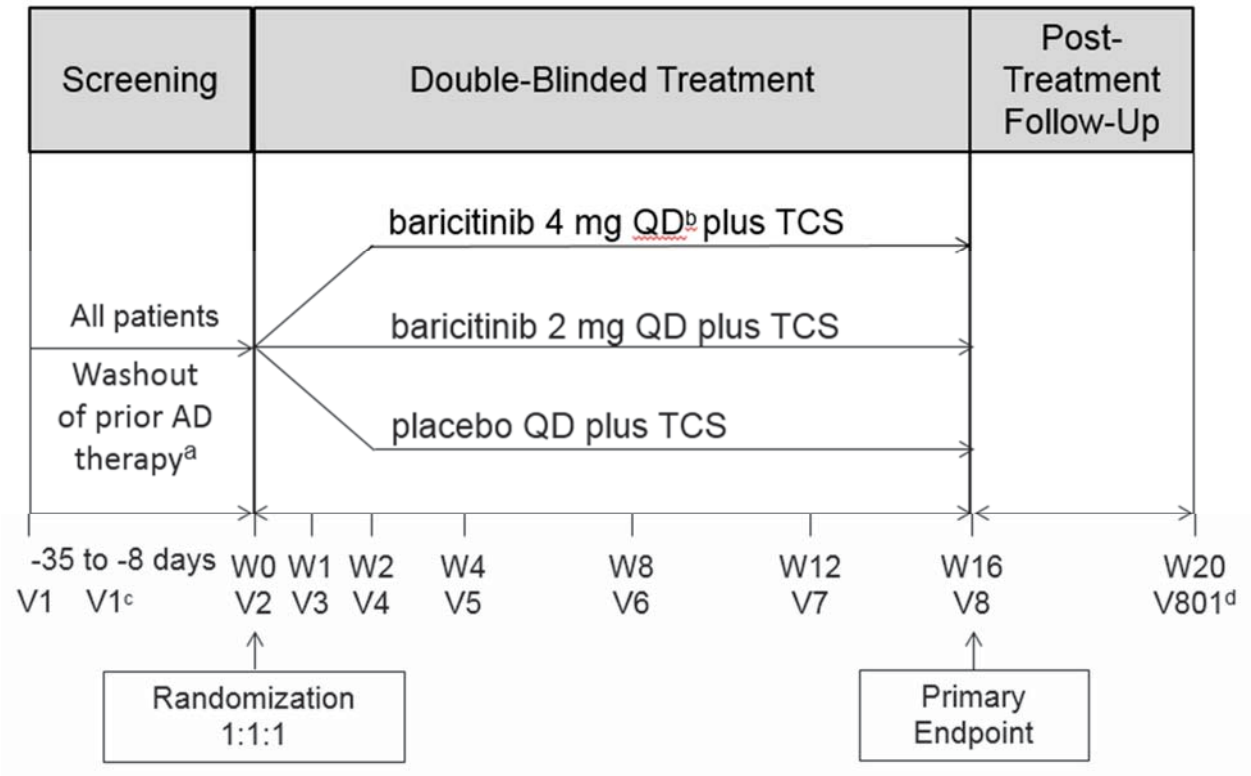
Patients who complete the study through Visit 8 (Week 16) and do not enter the long-term extension study will have a post-treatment follow-up visit (Visit 801) approximately 28 days after the last dose of investigational product.

Patients who have received at least 1 dose of investigational product and discontinue early from the study must have an ETV, and return for the post-treatment safety follow-up visit (Visit 801) approximately 28 days after the last dose of investigational product.

Patients who have discontinued investigational product, but remain in the study for more than 28 days without investigational product will have an ETV if they chose to discontinue early; however, a separate follow-up visit (V801) is not required.

Patients should not initiate new systemic AD treatment during this period. However, if patients or investigators must initiate treatment, patients should complete an unscheduled visit prior to the first dose of the new therapy.

Figure JAIY.5.1 illustrates the study design. The 3 dosing regimens are described in Section 7.1. The blinding procedure is described in Section 7.3.



Abbreviations: AD = atopic dermatitis; eGFR = estimated glomerular filtration rate; PPD = purified protein derivative; QD = once daily; TCS = topical corticosteroids; V = visit; W = week.

[a] Applicable to patients taking topical treatments (excluding emollients) or systemic treatments for AD at the time of screening.

[b] For patients randomized to the 4-mg QD dose who have renal impairment (defined as eGFR <60 mL/min/1.73 m^2), the baricitinib dose will be 2-mg QD.

[c] Patients for whom PPD skin test for the evaluation of tuberculosis infection was performed at V1 must return and PPD test must be read 48-72 hours after Visit 1 (post-PPD).

[d] Occurs approximately 28 days after the last dose of IP. Not required for those patients entering the long-term extension study JAHN.

Figure JAIY.5.1. Illustration of study design for Clinical Protocol I4V-MC-JAIY.

5.2. Number of Participants

Approximately 300 participants will be enrolled; approximately 420 patients will be screened to achieve this enrollment.

5.3. End of Study Definition

End of the study is the date of the last visit or last scheduled procedure shown in the Schedule of Activities (Section 2) for the last patient.

5.4. Scientific Rationale for Study Design

This study will enroll moderate to severe AD patients with a history of inadequate response to existing topical therapies for whom a systemic treatment such as baricitinib may therefore be appropriate.

Topical corticosteroids are the first-line anti-inflammatory treatment, even for patients treated with systemic treatments. For this reason, this study will assess the efficacy of baricitinib in combination with background mild-to-moderate potency TCS, used as determined appropriate by the investigator.

During the screening period (Period 1), a washout of systemic and topical treatments for AD was incorporated prior to randomization to minimize confounding effects of patients receiving a wide range of different background treatments prior to study entry as well as potential safety issues related to the use of other therapies, including rebound effects after discontinuation of systemic therapies. The double-blind, placebo-controlled treatment period (Period 2) is designed to minimize bias in the evaluation of the efficacy and safety of 2 baricitinib doses, relative to placebo, through 16 weeks of treatment.

In consideration of the disease severity, all patients in Study JAIY are eligible for rescue to higher potency TCS. Investigators are allowed to rescue patients who are experiencing unacceptable or worsening symptoms of AD. Once rescue medication is used, the patient will be determined to be a nonresponder (Section 7.7.4).

Investigator's Global Assessments are commonly used in clinical trials, both for qualifying patients for enrollment and for evaluating treatment efficacy (Langley et al. 2015; Futamura et al. 2016). There is no single "gold standard" disease severity scale for AD; however, IGA scales provide clinically meaningful measures to patients and investigators that are easily described and that correspond to disease severity categories (e.g., moderate to severe). The scale that will be used in this study, the validated Investigator's Global Assessment of Atopic Dermatitis (vIGA-AD, referred to throughout the protocol as IGA), has been developed internally and assesses AD severity using a 5-point scale.

The 16-week efficacy endpoint was chosen because it is likely that a robust clinical effect will be observed with baricitinib within this timeframe based on the Phase 2 study results in AD and from previous studies in another inflammatory skin condition.

The Post-Treatment Follow-Up Period (Period 3) is for safety monitoring after the patient has been off investigational product for approximately 28 days.

5.5. Justification for Dose

The doses proposed for AD Phase 3 studies are baricitinib 2-mg, and 4-mg QD. These doses were chosen primarily based on the recently completed Phase 2 AD study, JAHG, and are additionally supported by pharmacokinetics (PK), safety, and efficacy data for baricitinib in Phase 2 and Phase 3 RA studies and a Phase 2 psoriasis study.

In the Phase 2 Study JAHG, both the 2-mg and 4-mg doses showed benefit on the primary and major secondary endpoints (Eczema Area and Severity Index [EASI], IGA, SCORing Atopic Dermatitis [SCORAD], Patient-Oriented Eczema Measure [POEM], and Dermatology Life Quality Index [DLQI]) as compared to placebo, and both doses had an acceptable safety profile at Week 16 (Guttman-Yassky et al. 2018). However, the 4-mg dose appeared to demonstrate a more rapid benefit (at 4 weeks) on the more stringent endpoints (improvement of at least 75% in EASI score [EASI75], improvement of at least 90% in EASI score [EASI90], and IGA 0 or 1) compared to 2-mg dose particularly in the subgroup of patients with baseline EASI scores ≥16. The 4-mg dose resulted in statistically significant improvement in these endpoints at Week 4 and this level of response was maintained through Week 16. A similar trend between the baricitinib 4-mg and 2-mg doses was observed in patients with RA. Although in Study JAHG, the 4-mg dose seemed to perform better than the 2-mg dose on more stringent endpoints, on other endpoints, including an improvement of at least 50% in EASI score [EASI50], and EASI change from baseline, 2-mg and 4-mg doses showed similar efficacy compared to placebo. Therefore, both doses will be tested in Study JAIY.

5.5.1. Dose Adjustment for Renal Impairment

Baricitinib exposure increases with decreased renal function. Based on PK simulations of baricitinib exposures for the mild and moderate categories of renal function (stratified as estimated glomerular filtration rate [eGFR] 60 to <90 mL/min/1.73 m^2 and eGFR 30 to <60 mL/min/1.73 m^2, respectively), dose adjustment is not required for patients with eGFR ≥60 mL/min/1.73 m^2. Patients with eGFR <60 mL/min/1.73 m^2 who are randomized to the 4-mg dose will receive a dose of 2-mg QD, which will ensure that exposures do not exceed those of the 4-mg QD dose in patients with eGFR ≥60 mL/min/1.73 m^2. For patients randomized to the 2-mg dose, there will be no dose adjustment based on renal function. The dose adjustment for renal impairment will be managed by interactive web-response system (IWRS) to ensure maintenance of the treatment blind. This study will not enroll patients with screening eGFR <40 mL/min/1.73 m^2. See Section 8.1.1 for eGFR thresholds that trigger interruption of investigational product.

The procedure of dose adjustment based on renal function (eGFR) during the study is detailed in Section 7.2.2.

6. Study Population

Prospective approval of protocol deviations to recruitment and enrollment criteria, also known as protocol waivers or exemptions, are not permitted.

Study investigator(s) will review patient history and screening test results at Visit 1 and Visit 2 to determine if the patient meets all inclusion and none of the exclusion criteria to qualify for randomization in the study. All screening activities must be completed and reviewed before the patient is randomized.

6.1. Inclusion Criteria

Informed Consent

[1] are at least 18 years of age at the time of informed consent.

Note: Use local requirements to provide consent if the age of adulthood is defined as >18 years

[2] are able to read, understand, and give documented (electronic or paper signature) informed consent.

Type of Patient and Disease Characteristics

[3] have a diagnosis of AD at least 12 months prior to screening, as defined by the American Academy of Dermatology: Guidelines of care for the management of AD; Section 1. Diagnosis and assessment of atopic dermatitis (Appendix 7).

[4] have moderate to severe AD, including all of the following:

a. Eczema Area and Severity Index (EASI) score ≥16 at screening (Visit 1) and at randomization (Visit 2)

b. IGA score of ≥3 at screening (Visit 1) and at randomization (Visit 2)

c. ≥10% of BSA involvement at screening (Visit 1) and at randomization (Visit 2).

[5] have a documented history by a physician and/or investigator of inadequate response to existing topical medications within 6 months preceding screening as defined by at least 1 of the following:

a. inability to achieve good disease control defined as mild disease or better (e.g., IGA ≤2) after use of at least a moderate potency TCS for at least 4 weeks, or for the maximum duration recommended by the product prescribing information (e.g., 14 days for super-potent TCS), whichever is shorter. Topical corticosteroids may be used with or without TCNIs.

b. Patients who failed systemic therapies intended to treat AD within 6 months preceding screening, such as cyclosporine, methotrexate, azathioprine, and mycophenolate mofetil will also be considered as a surrogate for having inadequate response to topical therapy.

[6] agree to discontinue use of the following excluded medications/treatments for at least 4 weeks prior to randomization (Visit 2) and throughout the study:

a. oral systemic corticosteroids

b. systemic immunomodulators, including, but not limited to, cyclosporine, methotrexate, mycophenolate mofetil, and azathioprine

c. any other systemic therapy used to treat AD or symptoms of AD (approved or off-label use)

[7] agree to discontinue the use of following excluded medications for at least 2 weeks prior to randomization (Visit 2):

a. TCS or topical immune modulators (e.g., tacrolimus or pimecrolimus)

b. Topical phosphodiesterase type 4 (PDE-4) inhibitor (crisaborole)

c. sedating antihistamines, including, but not limited to, alimemazine, chlorphenamine, clemastine, cyproheptadine, diphenhydramine, hydroxyzine, ketotifen, and promethazine

Note: Patients may use newer, less sedating antihistamines (Section 7.7.1) for the treatment of allergic conditions other than AD. Use of antihistamines for the treatment of itch is not allowed during the study.

d. phototherapy, includes therapeutic phototherapy (psoralen plus ultraviolet A, ultraviolet B), excimer laser as well as self-treatment with tanning beds.

[8] have applied emollients daily for at least 14 days prior to randomization and agree to use emollient daily throughout the treatment period.

[9] Patients who are receiving chronic treatments to improve sleep should be on a stable dose for at least 2 weeks prior to screening as determined by the investigator. Sedating antihistamines (see above) are not permitted.

Patient Characteristics

[10] Male or nonpregnant, nonbreastfeeding female patients

Patients of child-bearing potential who are abstinent (if this is complete abstinence, as their preferred and usual lifestyle) or in a same-sex relationship (as part of their preferred and usual lifestyle) must agree to either remain abstinent or stay in a same-sex relationship without sexual relationships with the opposite sex.

Total abstinence is defined as refraining from intercourse during the entirety of the study and for at least 1 week following the last dose of investigational product. Periodic abstinence such as calendar, ovulation, symptothermal, postovulation methods and withdrawal are not acceptable methods of contraception.

Otherwise, patients and their partners of child-bearing potential must agree to use 2 effective methods of contraception, where at least 1 form is highly effective for the entirety of the study and for at least 1 week following the last dose of investigational product.

The following contraception methods are considered acceptable (the patient, and their partner, should choose 2, and 1 must be highly effective [defined as <1% failure rate per year when used consistently and correctly]):

- Highly effective birth control methods:
  - Combined (estrogen- and progestogen-containing) hormonal contraception associated with inhibition of ovulation: oral, intravaginal, or transdermal
  - Progestogen-only hormonal contraception associated with inhibition of ovulation: oral, intravaginal, or implantable
  - Intrauterine device/intrauterine hormone-releasing system
  - Vasectomized partner (with appropriate post-vasectomy documentation of the absence of sperm in the ejaculate).
- Effective birth control methods:
  - Male or female condom with spermicide. It should be noted that the use of male and female condoms as a double barrier method is not considered acceptable due to the high failure rate when these methods are combined.
  - Diaphragm with spermicide
  - Cervical sponge
  - Cervical cap with spermicide
  - Oral contraceptives that do not inhibit ovulation

  Note: When local guidelines concerning highly effective or effective methods of birth control differ from the above, the local guidelines must be followed.

a. Females of non–child-bearing potential are not required to use birth control and they are defined as:

   - women ≥60 years of age or women who are congenitally sterile, or
   - women ≥40 and <60 years of age who have had a cessation of menses for ≥12 months and a follicle-stimulating hormone (FSH) test confirming nonchildbearing potential (≥40 mIU/mL or ≥40 IU/L), or women who are surgically sterile (i.e., have had a hysterectomy or bilateral oophorectomy or tubal ligation).

6.2. Exclusion Criteria

Patients will be excluded from study enrollment if they meet any of the following criteria:

Medical Conditions Related to Atopic Dermatitis

[11] are currently experiencing or have a history of other concomitant skin conditions (e.g., psoriasis or lupus erythematosus) that would interfere with evaluations of the effect of study medication on AD.

[12] have had an important side effect to TCS (e.g., intolerance to treatment, hypersensitivity reactions, significant skin atrophy, and systemic effects), as assessed by the investigator or treating physician that would prevent further use.

[13] patients who, in the opinion of the investigator, are currently experiencing or have a history of erythrodermic, refractory, or unstable skin disease that requires frequent hospitalizations and/or intravenous treatment for skin infections that may interfere with participation in the study.

[14] a history of eczema herpeticum within 12 months prior to screening.

[15] a history of 2 or more episodes of eczema herpeticum in the past.

[16] patients who are currently experiencing a skin infection that requires treatment, or is currently being treated, with topical or systemic antibiotics.

Note: Patients may not be rescreened until at least 4 weeks after the date of their previous screen failure and at least 2 weeks after resolution of the infection.

[17] have any serious concomitant illness that is anticipated to require the use of systemic corticosteroids or otherwise interfere with study participation or require active frequent monitoring (e.g., unstable chronic asthma).

[18] have been treated with the following therapies:

a. monoclonal antibody (e.g., ustekinumab, omalizumab, and dupilumab) or fusion proteins that target inflammatory pathways (e.g., etanercept) for less than 5 half-lives prior to randomization.

b. received prior treatment with any oral JAK inhibitor (e.g., tofacitinib and ruxolitinib) <4 weeks prior to randomization

c. received any parenteral corticosteroid administered by intramuscular or intravenous injection within 6 weeks prior to planned randomization (Visit 2) or are anticipated to require parenteral injection of corticosteroids during the study.

d. have had an intra-articular corticosteroid injection within 6 weeks prior to planned randomization (Visit 2).

Note: Intranasal or inhaled steroid use is allowed during the trial.

e. probenecid at the time of randomization (Visit 2) that cannot be discontinued for the duration of the study

Medical Conditions in General

[19] are largely or wholly incapacitated permitting little or no self-care, such as being bed-ridden.

[20] have uncontrolled arterial hypertension characterized by a repeated systolic blood pressure >160 mm Hg or diastolic blood pressure >100 mm Hg in a seated position. Reassessment of blood pressure during the screening period is allowed.

[21] have had any major surgery within 8 weeks prior to screening or will require major surgery during the study that in the opinion of the investigator in consultation with Lilly or its designee, would pose an unacceptable risk to the patient if participating in the trial.

[22] are immunocompromised and, in the opinion of the investigator, at an unacceptable risk for participating in the study.

[23] have experienced any of the following within 12 weeks of screening: VTE, myocardial infarction (MI), unstable ischemic heart disease, stroke, or New York Heart Association Stage III/IV heart failure.

[24] have a history of recurrent (≥2) VTE or are considered at high risk of VTE as deemed by the investigator.

[25] have a history or presence of cardiovascular, respiratory, hepatic, gastrointestinal, endocrine, hematological, neurological, or neuropsychiatric disorders or any other serious and/or unstable illness that in the opinion of the investigator, could constitute an unacceptable risk when taking investigational product or interfere with the interpretation of data.

[26] have a history of lymphoproliferative disease; or have signs or symptoms suggestive of possible lymphoproliferative disease, including lymphadenopathy or splenomegaly; or have active primary or recurrent malignant disease; or have been in remission from clinically significant malignancy for less than 5 years.

a. Patients with cervical carcinoma in situ that has been appropriately treated with no evidence of recurrence or metastatic disease for at least 3 years may participate in the study.

b. Patients with basal cell or squamous epithelial skin cancers that have been appropriately treated with no evidence of recurrence for at least 3 years may participate in the study.

[27] have a current or recent clinically serious viral, bacterial, fungal, or parasitic infection, including but not limited to the following:

Note: A recent viral upper respiratory tract infection or uncomplicated urinary tract infection should not be considered clinically serious.

a. symptomatic herpes zoster infection within 12 weeks prior to screening.

b. a history of disseminated/complicated herpes zoster (e.g., multidermatomal involvement, ophthalmic zoster, central nervous system involvement, or postherpetic neuralgia).

c. symptomatic herpes simplex at the time of randomization.

d. active or chronic viral infection from hepatitis B virus (HBV), hepatitis C virus (HCV), or human immunodeficiency virus (HIV).

e. household contact with a person with active tuberculosis (TB) and did not receive appropriate and documented prophylaxis for TB.

f. evidence of active TB or have previously had evidence of active TB and did not receive appropriate and documented treatment.

g. clinically serious infection or received intravenous antibiotics for an infection, within the past 4 weeks of randomization.

h. any other active or recent infection within 4 weeks of randomization that, in the opinion of the investigator, would pose an unacceptable risk to the patient if participating in the study.

[28] have been exposed to a live vaccine within 12 weeks prior to planned randomization or are expected to need/receive a live vaccine during the course of the study (with the exception of herpes zoster vaccination).

Note: Patients eligible for herpes zoster vaccine, who have not received it prior to screening will be encouraged (per local guidelines) to do so prior to randomization; vaccination with the live herpes zoster vaccine can occur during the screening period but must take place >4 weeks prior to randomization and start of IP.

Vaccination with the non-live herpes zoster vaccine requires at least 2 injections administered 8 weeks apart and, therefore, cannot be completed during the 5 week screening period. Patients who have initiated vaccination with non-live herpes zoster vaccine before the trial should plan to receive the second dose at least 4 weeks prior to randomization.

[29] have a history of chronic alcohol abuse, intravenous drug abuse, or other illicit drug abuse within the 2 years prior to screening.

[30] presence of significant uncontrolled neuropsychiatric disorder, are clinically judged by the investigator to be at risk for suicide, or have a “yes” answer to any of the following:

a. question 4 (Active Suicidal Ideation with Some Intent to Act, Without Specific Plan) on the “Suicidal Ideation” portion of the Columbia Suicide Severity Rating Scale (C-SSRS) or

b. question 5 (Active Suicidal Ideation with Specific Plan and Intent) on the "Suicidal Ideation" portion of the C-SSRS or

c. any of the suicide-related behaviors (actual attempt, interrupted attempt, aborted attempt, and preparatory act or behavior) on the "Suicidal Behavior" portion of the C-SSRS;

and the ideation or behavior occurred within 2 months prior to Visit 1.

Note: a patient does not necessarily have to be excluded if they have self-injurious behavior that would be classified as nonsuicidal self-injurious behavior. If this situation arises, the subject should be referred to a psychiatrist or appropriately trained professional as indicated.

[31] have donated more than a single unit of blood within 4 weeks prior to screening or intend to donate blood during the course of the study.

Other Exclusions

[32] are unable or unwilling to make themselves available for the duration of the study and/or are unwilling to follow study restrictions/procedures.

[33] are currently enrolled in any other clinical trial involving an investigational product or any other type of medical research judged not to be scientifically or medically compatible with this study.

[34] have participated within the last 30 days in a clinical study involving an investigational product. If the previous investigational product has a long half-life (2 weeks or longer), at least 3 months or 5 half-lives (whichever is longer) should have passed.

[35] have previously been randomized in this study or any other study investigating baricitinib, or who have experienced hypersensitivity to the active substance or to any of the excipients.

[36] are investigator site personnel directly affiliated with this study and/or their immediate families. Immediate family is defined as a spouse, parent, child, or sibling, whether biological or legally adopted.

[37] are Lilly or Incyte employees or their designee.

Diagnostic Assessments

[38] have screening electrocardiogram (ECG) abnormalities that, in the opinion of the investigator, are clinically significant and indicate an unacceptable risk for the patient's participation in the study.

[39] have evidence of active TB or latent TB

a. have evidence of active TB, defined in this study as the following:

- documented by a positive PPD test (≥5 mm induration between approximately 48 and 72 hours after application, regardless of vaccination history), medical history, clinical features, and abnormal chest x-ray at screening.
- The QuantiFERON®-TB Gold test or T-SPOT®.TB test (as available and if compliant with local TB guidelines) may be used instead of the PPD test. Patients are excluded from the study if the test is not negative and there is clinical evidence of active TB.

Exception: Patients with a history of active TB who have documented evidence of appropriate treatment, have no history of re-exposure since their treatment was completed, and have a screening chest x-ray with no evidence of active TB may be enrolled if other entry criteria are met. Such patients would not be required to undergo the protocol-specific TB testing for PPD, QuantiFERON®-TB Gold test, or T-SPOT® TB test but must have a chest x-ray at screening.

b. have evidence of untreated/inadequately or inappropriately treated latent TB, defined in this study as the following:

- documented to have a positive PPD test (≥5 mm induration between approximately 48 and 72 hours after application, regardless of vaccination history), no clinical features consistent with active TB, and a chest x-ray with no evidence of active TB at screening; or
- PPD test is positive and the patient has no medical history or chest x-ray findings consistent with active TB, the patient may have a QuantiFERON®-TB Gold test or T-SPOT® TB test (as available and if compliant with local TB guidelines). If the test results are not negative, the patient will be considered to have latent TB (for purposes of this study); or
- QuantiFERON®-TB Gold test or T-SPOT® TB test (as available and if compliant with local TB guidelines) may be used instead of the PPD test. If the test results are positive, the patient will be considered to have latent TB. If the test is not negative, the test may be repeated once within approximately 2 weeks of the initial value. If the repeat test results are again not negative, the patient will be considered to have latent TB (for purposes of this study). Patients who have an indeterminate QuantiFERON®-TB Gold test (not negative), may either repeat the QuantiFERON®-TB Gold test or have T-SPOT TB test. Purified protein derivative testing after an indeterminate QuantiFERON®-TB Gold test is not allowed.

Exception: A patient who has evidence of latent TB may be enrolled if he or she completes at least 4 weeks of appropriate treatment prior to randomization and agrees to complete the remainder of treatment while in the trial.

Exception: Patients with a history of latent TB who have documented evidence of appropriate treatment, have no history of re-exposure since their treatment was completed, and have a screening chest x-ray with no evidence of active TB may be enrolled if other entry criteria are met. Such patients would not be required to undergo the protocol-specific TB testing for PPD, QuantiFERON®-TB Gold test, or T-SPOT® TB test but must have a chest x-ray at screening.

[40] have a positive test for HBV infection defined as:

a. positive for hepatitis B surface antigen (HBsAg), or

b. positive for hepatitis B core antibody (HBcAb) and positive hepatitis B virus deoxyribonucleic acid (HBV DNA).

Note: Patients who are HBcAb positive and HBV DNA negative may be enrolled in the study. Patients who meet these criteria at screening will be identified by the central laboratory and monitored during the study.

[41] have HCV infection (positive for anti-hepatitis C antibody with confirmed presence of HCV ribonucleic acid (RNA)

Note: Patients who have documented anti-HCV treatment for a past HCV infection AND are HCV RNA negative may be enrolled in the study.

[42] have evidence of HIV infection and/or positive HIV antibodies.

[43] have screening laboratory test values, including thyroid-stimulating hormone (TSH), outside the reference range for the population or investigative site that, in the opinion of the investigator, pose an unacceptable risk for the patient's participation in the study.

Note: Patients who are receiving thyroxine as replacement therapy may participate in the study, provided stable therapy has been administered for ≥12 weeks and TSH is within the laboratory's reference range. Patients who are receiving stable thyroxine replacement therapy who have TSH marginally outside the laboratory's normal reference range may participate if the treating physician has documented that the thyroxine replacement therapy is adequate for the patient.

[44] have any of the following specific abnormalities on screening laboratory tests:

a. AST or ALT ≥2x upper limit of normal (ULN)

b. alkaline phosphatase (ALP) ≥2x ULN

c. total bilirubin ≥1.5x ULN

d. hemoglobin <10.0 g/dL (100.0 g/L)

e. total white blood cell count <2500 cells/μL (<2.50x10^3/μL or <2.50 GI/L)

f. neutropenia (absolute neutrophil count [ANC] <1200 cells/μL) (<1.20x10^3/μL or <1.20 GI/L)

g. lymphopenia (lymphocyte count <750 cells/μL) (<0.75x10^3/μL or <0.75 GI/L)

h. thrombocytopenia (platelets <100,000/μL) (<100x10^3/μL or <100 GI/L)

i. eGFR <40 mL/min/1.73 m^2 (Chronic Kidney Disease Epidemiology Collaboration equation [CKD-EPI] Creatinine 2009 equation).

Note: For cases with any of the aforementioned laboratory abnormalities (Exclusion Criteria [43] and [44]), the tests may be repeated during screening, and values resulting from repeat testing may be accepted for enrollment eligibility if they meet the eligibility criterion.

6.3. Lifestyle Restrictions

Not applicable.

6.4. Screen Failures

Patients who are entered into the study but do not meet the eligibility criteria for participation in this study (screen failure) may be rescreened a maximum of 2 times. If patients are rescreened, rescreening cannot occur until at least 4 weeks after the date of their previous screen failure. When rescreening is performed, the individual must sign a new ICF and will be assigned a new identification number. Additionally, all necessary screening procedures must be conducted at rescreen to ensure all eligibility criteria are met.

7. Treatments

7.1. Treatments Administered

This study involves a comparison of placebo, baricitinib 2-mg, and baricitinib 4-mg administered orally once a day. Table JAIY.3 shows the treatment regimens.

Table JAIY.3. Treatment Regimens

Regimen	Investigational Product Supplied	Dose
Baricitinib 4-mg QD[a]	Baricitinib 4-mg tablets Placebo to match 2-mg tablets	2 tablets per day
Baricitinib 2-mg QD	Baricitinib 2-mg tablets Placebo to match 4-mg tablets	2 tablets per day
Placebo QD	Placebo to match 4-mg tablets Placebo to match 2-mg tablets	2 tablets per day

Abbreviation: QD = once daily.

[a] The baricitinib dose for patients randomized to the 4-mg QD treatment group who have renal impairment (defined as eGFR <60 mL/min/1.73 m^2) will be 2-mg QD.

The investigator or his or her designee is responsible for the following:

- explaining the correct use of the investigational agent(s) to the patient
- verifying that instructions are followed properly
- maintaining accurate records of investigational product dispensing and collection
- at the end of the study, returning all unused medication to Lilly, or its designee, unless the sponsor and sites have agreed all unused medication is to be destroyed by the site, as allowed by local law

7.1.1. Packaging and Labeling

The sponsor (or its designee) will provide the following investigational products:

- tablets containing 4-mg of baricitinib
- tablets containing 2-mg of baricitinib
- placebo tablets to match baricitinib 4-mg tablets and 2-mg tablets

Patients are required to take 2 tablets daily from packages assigned by the IWRS.

Clinical trial materials will be labeled according to the country's regulatory requirements.

7.2. Method of Treatment Assignment

Patients who meet all criteria for enrollment will be randomized in a 1:1:1 ratio (placebo: baricitinib 2-mg: baricitinib 4-mg) to double-blind treatment at Visit 2 (Week 0). Randomization will be stratified by geographic region (Europe, Japan, rest-of-world) and disease severity at baseline (IGA 3 vs. 4). Assignment to treatment groups will be determined by a computer-generated random sequence using an IWRS. The IWRS will be used to assign

packages containing double-blind investigational product tablets to each patient according to the study schedule of activities. Site personnel will confirm that they have located the correct packages by entering a confirmation number found on the packages into the IWRS.

7.2.1. Selection and Timing of Doses

The investigational product should be taken once daily without regard to food and if possible, at approximately the same time every day, usually at the start of the patient's day, to aid patient compliance.

7.2.2. Dose Adjustment for Renal Impairment

The rationale of dose adjustment for patients with documented renal impairment (defined as screening eGFR ≥40 to <60 mL/min/1.73 m^2) is detailed in Section 5.5.1.

The dose adjustment for renal impairment will be managed by IWRS to ensure maintenance of the treatment blind. The eGFR value from the screening visit (Visit 1) will be entered into IWRS at Visit 2, and IWRS will assign the treatment doses accordingly.

Patients with documented renal impairment (defined as screening eGFR ≥40 to <60 mL/min/1.73 m^2), who are randomized to the 4-mg active treatment arm will receive a dose of 2-mg QD by the IWRS. For patients randomized to the 2-mg dose, there will be no dose adjustment based on renal function.

No dose adjustment will be made for patients with screening eGFR ≥60 mL/min/1.73 m^2. These patients who are randomized to active treatment will receive their assigned dose, either baricitinib 4-mg or 2-mg, respectively.

During the study, for patients with documented renal impairment when the subsequent eGFR falls <30 mL/min/1.73 m^2, investigational product will be withheld until their eGFR becomes ≥40 mL/min/1.73 m^2, whereupon the investigational product dosing may resume. For patients with screening eGFR ≥60 mL/min/1.73 m^2, when the subsequent eGFR falls to <40 mL/min/1.73 m^2, investigational product will be withheld until their eGFR becomes ≥50 mL/min/1.73 m^2, whereupon the investigational product dosing may resume (Section 8.1.1).

7.3. Blinding

This is a double-blind study. To preserve the blinding of the study, a minimum number of Lilly personnel will see the randomization table and treatment assignments before the study is complete. All study assessments will be performed by study personnel who are blinded to the patient's treatment group. Except in clinical circumstances where unblinding is required, the patients, investigators, Lilly study team, and any personnel interacting directly with patients or investigative sites will remain blinded to baricitinib and placebo assignment until after completion of the Double-Blinded Treatment Period. It is expected that the need for unblinding a patient's treatment prior to completion of the Double-Blinded Treatment Period will be extremely rare. Every effort should be made to preserve the blind unless there is a compelling reason that knowledge of the specific treatment would alter the medical care of the patient. In case of an emergency, the investigator has the sole responsibility for determining if unblinding of

a patient's treatment assignment is warranted for medical management of the event. Patient safety must always be the first consideration in making such a determination. Where feasible and when timing of the emergent situation permits, the investigator should attempt to contact the Lilly medical monitor before unblinding a subject's treatment assignment. If a patient's treatment assignment is unblinded, Lilly must be notified immediately. If the investigator decides that unblinding is warranted, it is the responsibility of the investigator to promptly document the decision and rationale and notify Lilly as soon as possible.

Emergency unblinding for AEs may be performed through the IWRS. This option may be used ONLY if the patient's well-being requires knowledge of the patient's treatment assignment. All unblinding events are recorded and reported by the IWRS. If an investigator, site personnel performing assessments, or patient is unblinded, the patient must be discontinued from the study. In cases where there are ethical reasons to have the patient remain in the study, the investigator must obtain specific approval from a Lilly clinical research physician for the patient to continue in the study.

Processes to maintain blinding during the interim analysis conducted by the DMC are described in Section 10.3.7.1.

7.4. Dosage Modification

Not applicable.

7.5. Preparation/Handling/Storage/Accountability

All investigational product (used and partially used) will be returned to the sponsor or destroyed at site level with the sponsor's written approval. In some cases, sites may destroy the material if, during the investigative site selection, the evaluator has verified and documented that the site has appropriate facilities and written procedures to dispose of clinical trial materials.

Follow storage and handling instructions on the investigational product packaging.

7.6. Treatment Compliance

Patient compliance with study medication will be assessed at each visit during the treatment period (Visit 3 through Visit 8) by counting returned tablets.

A patient will be considered significantly noncompliant if he or she misses more than 20% of the prescribed doses of investigational product during the study, unless the patient's investigational product is withheld by the investigator for safety reasons. Similarly, a patient will be considered significantly noncompliant if he or she is judged by the investigator to have intentionally or repeatedly taken 20% more than the prescribed amount of medication during the study.

Patients will be counseled by study staff on the importance of taking the investigational product as prescribed, as appropriate.

Patients' compliance will be further defined in the statistical analysis plan (SAP).

7.7. Concomitant Therapy

All concomitant medication, whether prescription or over the counter, used at baseline and/or during the course of the study, must be recorded on the Concomitant Medication electronic case report form (eCRF). Patients will be instructed to consult the investigator or other appropriate study personnel at the site before taking any new medications or supplements during the study. For AD therapies permitted as part of rescue therapy, see Section 7.7.4.

7.7.1. Permitted Medications and Procedures

Treatment with concomitant AD therapies during the study is permitted only as described below.

- Daily use of emollients is required as background treatment. Moisturizers with additives containing pharmacological agents with antipruritic or antiseptic properties are not permitted. If daily applications are missed, it will not be considered a protocol violation
  - Patients should not apply emollients on the day of their study visit prior to the procedures to allow adequate assessment of skin dryness
- Background TCS therapy with moderate-potency and/or low-potency TCS (e.g., triamcinolone 0.1% cream and/or hydrocortisone 2.5% ointment) are to be used on active lesions, as described in Section 7.7.2
- Topical calcineurin inhibitors (e.g., tacrolimus and pimecrolimus), or topical PDE-4 inhibitor (i.e., crisaborole, where approved) are permitted in place of TCS on areas where application of TCS is considered inappropriate by the investigator; use should be limited to problem areas (e.g., face, neck, skin folds, genital areas, etc.) as described in Section 7.7.2

In addition, the following therapies are permitted during the study:

- For those patients on stable dosing of prescription sleep medications at entry, downward dose adjustments or discontinuation of treatment may occur during the study
- Nonsedating antihistamines including, but not limited to, acrivastine, bilastine, cetirizine, desloratadin, fexofenadine, levocetirizine, loratadine, mizolastine, and rupatadine are allowed for treatment of allergic conditions other than AD. Use of antihistamines for treatment of itch is not allowed during the study
- Single intra-articular or soft tissue (bursa, tendons, and ligaments) corticosteroid injection is allowed during the 16-week double-blind, placebo-controlled period
- Intranasal or inhaled steroid use is allowed
- Topical anesthetics and topical and systemic anti-infective medications are allowed
- Nonlive vaccinations are allowed; however, vaccination while receiving baricitinib may reduce efficacy of the vaccine
- ophthalmic drugs containing antihistamines, corticosteroids or other immunosuppressants are allowed

Any changes of these concomitant medications must be recorded in the Concomitant Therapy of Special Interest eCRF.

Treatment with concomitant therapies for other medical conditions such as diabetes and hypertension is permitted during the study.

7.7.2. Use of Topical Corticosteroids

A washout period of 14 days is required for all TCS prior to randomization at Visit 2.

At baseline (Week 0, Visit 2), patients will receive triamcinolone 0.1% cream (or equivalent-potency TCS) and hydrocortisone 2.5% ointment (or equivalent-potency TCS). See "Choice of Background Topical Corticosteroid" below. Triamcinolone 0.1% cream (moderate-potency TCS) should be applied at least once daily to affected areas until lesions are under control (clear or almost clear). Patients should then switch to hydrocortisone 2.5% ointment (low-potency TCS) and treat previously affected areas once-daily for 7 days and then stop. Hydrocortisone 2.5% ointment (low-potency TCS) may also be used to replace triamcinolone 0.1% cream (moderate-potency TCS) on areas of thin skin (face, neck, folds, and genital areas) and areas with skin atrophy.

If lesions reappear during the course of the study, the patients should resume the once-daily applications of triamcinolone 0.1% cream (moderate-potency TCS) or hydrocortisone 2.5% ointment (low-potency TCS) as described above.

Patients whose lesions persist or worsen despite the use of emollients and low- and/or moderate-potency TCS and/or patients who require daily applications on large surfaces may be considered for topical rescue with high- or ultra-high-potency TCS (Section 7.7.4 for details).

On the days of study visits, topical therapy including TCS should not be applied before the patient has undergone all study procedures and clinical evaluations in order to allow adequate assessment of skin dryness. Inability to follow this guidance for use of TCS will not be considered a protocol violation.

Choice of Background Topical Corticosteroid

Where possible, triamcinolone cream 0.1% and/or hydrocortisone 2.5% ointment will be supplied by the sponsor for use as background TCS. In the event of these specific TCS being unavailable, an alternate, equivalent-potency TCS may be provided by the sponsor (see below). Topical corticosteroid use, when supplied by the sponsor, should be recorded via weight of returned tubes as indicated in the Schedule of Activities (Section 2). In the event that the sponsor is unable to supply TCS, commercially available triamcinolone 0.1% cream and/or hydrocortisone 2.5% ointment may be supplied by the sites.

- Where providing triamcinolone 0.1% cream and/or hydrocortisone 2.5% ointment is not possible, an equivalent-potency TCS cream and/or ointment that is in line with local practices can be supplied by the sites. Refer to Appendix 8 for guidance on potency equivalence.

- Where possible, TCS use when supplied by the site should also be recorded via weight of returned tubes, as indicated in the Schedule of Activities (Section 2); however, where this is not practical, this information does not need to be recorded, and will not be considered a protocol violation.
- If the TCS supplied by the sponsor is not considered suitable for an individual patient, an equivalent-potency TCS cream and/or ointment that is in line with local practices can be supplied by the sites. Refer to Appendix 8 for guidance on potency equivalence.

Choice of High- and Ultra-High-Potency Topical Corticosteroids for Rescue

The use and choice of specific high- or ultra-high-potency TCS for rescue is at the discretion of the investigator and it will not be provided by the sponsor. The weights of returned tubes of the high- and ultra-high-potency TCS are not required.

Other Topical Treatments

Investigators may also select to use TCNIs and/or crisaborole in countries where approved, in place of TCS. If TCNIs or crisaborole are prescribed, use should be limited to problem areas (e.g., face, neck, skin folds, genital areas, etc.).

Use of all topical treatments for AD must be recorded in the CRF.

7.7.3. Prohibited Medications and Procedures

Prohibited Medications and Procedures Not Requiring Interruption of Investigational Product

The following therapies will not be allowed during the course of the study and, if taken by or administered to the patient, the prohibited therapy must be discontinued.

- High potency TCS (defined as any TCS with higher than moderate strength as defined in Appendix 8) except when given as rescue therapy as described in Section 7.7.4.
- topical antihistamines or sedating, systemic antihistamines including, but not limited to, alimemazine, chlorphenamine, clemastine, cyproheptadine, diphenhydramine, hydroxyzine, ketotifen, and promethazine
- Prescription allergen immunotherapy
- phototherapy including psoralen and ultraviolet A (PUVA), ultraviolet B, tanning booth and excimer laser
- bleach baths

Prohibited Medications Requiring Temporary Interruption of Investigational Product

The following therapies will not be allowed during the course of the study and, if taken by or administered to the patient, temporary interruption of investigational product is required.

- live vaccines (including Bacillus Calmette-Guérin [BCG] or herpes zoster), (see Exclusion Criterion [28])

  - For BCG vaccination, investigational product should be temporarily interrupted for 12 weeks.
  - For live herpes zoster vaccination, investigational product should be temporarily interrupted for 4 weeks.
- probenecid: if a patient is inadvertently started on probenecid, investigational product should be temporarily interrupted, and can be resumed after patient has discontinued probenecid. If a patient is not able to discontinue probenecid, then investigational product should be permanently discontinued

Prohibited Medications Requiring Permanent Discontinuation of Investigational Product

- systemic corticosteroids
- any systemic therapy, investigational or commercial (approved or off label use), used for the treatment of AD or symptoms of AD (except for antihistamines, as specified above)
- other JAK inhibitors (e.g., tofacitinib and ruxolitinib)
- systemic immunosuppressive/immunomodulatory substances, including, but not limited to, cyclosporine, methotrexate, mycophenolate mofetil, interferon γ, azathioprine, or biologic agents

Note: In the event that these prohibited medications were inadvertently used, agreement and documentation to continue investigational product must be sought from sponsor**.**

7.7.4. Rescue Therapy

For patients who are experiencing worsening and unacceptable symptoms of AD despite treatment with investigational product and moderate-potency TCS, rescue therapy with additional topical and systemic therapies is available starting after 2 weeks of treatment (Visit 4). Use of rescue medications should be limited to patients where control of symptoms cannot be achieved with increased emollient use and background TCS (low potency and moderate potency). Prior to rescue, control of AD symptoms should be attempted by avoiding exacerbating factors, intensifying emollient applications, and using only the permitted study treatments, including background TCS therapy (e.g., triamcinolone 0.1% cream and/or hydrocortisone 2.5% ointment) (Section 7.7.2).

Rescue with High- and Ultra-High-Potency TCS

Patients whose lesions persist or worsen despite the use of emollients and background TCS therapy (e.g., triamcinolone 0.1% cream and/or hydrocortisone 2.5% ointment) and/or patients who require prolonged applications of triamcinolone 0.1% cream (moderate-potency TCS) on large surfaces may be considered for rescue to high- or ultra-high-potency TCS (Appendix 8 for TCS potency).

High- or ultra-high-potency TCS may be used once daily for up to 14 consecutive days or less, or based on the maximum duration recommended in the prescribing information.

It is recommended that if a patient reaches "clear" to "almost clear" skin after topical rescue, then moderate-, high-, or ultra-high-potency TCS should be stopped, and low-potency TCS (e.g., hydrocortisone 2.5% ointment) should be used once daily for an additional 7 days, then stopped.

Patients rescued with high- or ultra-high-potency TCS will continue to take investigational product and use of topical rescue therapy will be documented in the eCRF.

Rescue with Systemic Therapies

If topical rescue therapy as described above fails to sufficiently control AD symptoms, then oral systemic medications may be used as rescue (e.g., corticosteroids, cyclosporine, and methotrexate); however, investigational product will be required to be permanently discontinued for the remainder of the 16-week study duration. If these medications are needed for other medical conditions (e.g., asthma flare), they will still be treated as rescue medications. These patients are still eligible to enter the long-term extension study (JAHN), if they complete the schedule of study visits through Visit 8 (Week 16) and are also able to complete a minimum 4-week washout from oral systemic rescue medications (which can occur during the screening period for Study JAHN).

Investigators should make every attempt to conduct efficacy and safety assessments immediately before administering any rescue treatment. An unscheduled visit can be used for this purpose if necessary.

7.8. Treatment after the End of the Study

7.8.1. Study Extensions

Patients who complete this study through Visit 8 may be eligible to participate in Study JAHN, if enrollment criteria for Study JAHN are met.

7.8.2. Continued Access

After the conclusion of the study, continued access to baricitinib will not be provided to patients who are not eligible for or who do not choose to participate in Study JAHN. Patients will be referred to their local treatment centers for AD therapy as clinically indicated.

8. Discontinuation Criteria

8.1. Discontinuation from Study Treatment

8.1.1. Temporary Interruption from Investigational Product

In some circumstances, it may be necessary to temporarily interrupt treatment as a result of AEs or abnormal laboratory values that may have an unclear relationship to investigational product. For example, investigational product should be temporarily interrupted if the patient experiences a cardiovascular AE considered to be related to the study treatment, is graded as moderate (Grade 2 according to Common Terminology Criteria for Adverse Events [CTCAE] Version 3.0), and that does not resolve promptly with supportive care. Except in cases of emergency, it is recommended that the investigator consult with Lilly (or its designee) before temporarily interrupting therapy for reasons other than those defined in Table JAIY.4.

For the abnormal laboratory findings and clinical events (regardless of relatedness) listed in Table JAIY.4, specific guidance is provided for temporarily interrupting treatment and when treatment may be restarted. Retest frequency and timing of follow-up laboratory tests to monitor the abnormal finding is at the discretion of the investigator. Investigational product that was temporarily interrupted because of an AE or abnormal laboratory value not specifically covered in Table JAIY.4 may be restarted at the discretion of the investigator. If laboratory abnormalities leading to temporary interruption of investigational product are identified from Visit 2 laboratory tests, investigational product must still be interrupted, even though the abnormal laboratory results are unrelated to investigational product.

Table JAIY.4. Criteria for Temporary Interruption of Investigational Product

Hold Investigational Product If the Following Laboratory Test Results or Clinical Events Occur:	Investigational Product May be Resumed When:
WBC count <2000 cells/µL (<2.00x10^3/µL or <2.00 GI/L)	WBC count ≥2500 cells/µL (≥2.50x10^3/µL or ≥2.50 GI/L)
ANC <1000 cells/µL (<1.00x10^3/µL or <1.00 GI/L)	ANC ≥1200 cells/µL (≥1.20x10^3/µL or ≥1.20 GI/L)
Lymphocyte count <500 cells/µL (<0.50x10^3/µL or <0.50 GI/L)	Lymphocyte count ≥750 cells/µL (≥0.75x10^3/µL or ≥0.75 GI/L)
Platelet count <75,000/µL (<75x10^3/µL or <75 GI/L)	Platelet count ≥100,000/µL (≥100x10^3/µL or ≥100 GI/L)
eGFR <40 mL/min/1.73 m^2 (from serum creatinine) for patients with screening eGFR ≥60 mL/min/1.73 m^2	eGFR ≥50 mL/min/1.73 m^2
eGFR <30 mL/min/1.73 m^2 (from serum creatinine) for patients with screening eGFR ≥40 to <60 mL/min/1.73 m^2	eGFR ≥40 mL/min/1.73 m^2
ALT or AST >5x ULN	ALT and AST return to <2x ULN, and IP is not considered to be the cause of enzyme elevation
Hemoglobin <8 g/dL (<80.0 g/L)	Hemoglobin ≥10 g/dL (≥100.0 g/L)
Symptomatic herpes zoster	All skin lesions have crusted and are resolving
Infection that, in the opinion of the investigator, merits the IP being interrupted	Resolution of infection
Clinical features of VTE (such as deep vein thrombosis or pulmonary embolism) are present[a]	After evaluation and institution of appropriate treatment of VTE[b]

Abbreviations: ALT = alanine aminotransferase; ANC = absolute neutrophil count; AST = aspartate aminotransferase; eGFR = estimated glomerular filtration rate; IP = investigational product; ULN = upper limit of normal; VTE = venous thromboembolic event; WBC = white blood cell.

[a] Evaluate promptly and institute appropriate treatment. Upon evaluation if VTE is ruled out and no other temporary or permanent discontinuation criteria are met, then IP may be resumed.

[b] After evaluation and institution of treatment if the investigator deems that the patient is still at significant risk, or if this would constitute a second VTE for the patient, then IP should be discontinued permanently.

Although temporary interruption of investigational product is not a requirement at times of increased potential risk of VTE (e.g., surgery, significant air travel, or other situations involving prolonged immobilization) we recommend following appropriate VTE prophylaxis guidelines to help manage the VTE risk under these circumstances.

For specific guidance on temporary interruption of investigational product after use of a prohibited medication, please refer to Section 7.7.3 (Prohibited Medications and Procedures).

Lastly, investigational product should be temporarily interrupted for suicidal ideation or any suicide-related behaviors as assessed by the following patient responses on the C-SSRS:

- A "yes" answer to Question 4 (Active Suicidal Ideation with Some Intent to Act, Without Specific Plan) **or**
- A "yes" answer to Question 5 (Active Suicidal Ideation with Specific Plan and Intent) on the "Suicidal Ideation" portion of the C-SSRS **or**
- A "yes" answer to any of the suicide-related behaviors (actual attempt, interrupted attempt, aborted attempt, preparatory act or behavior) on the "Suicidal Behavior" portion of the C-SSRS

NOTE: Prior to resumption of investigational product, it is recommended that a patient be assessed by a psychiatrist or appropriately trained professional to assist in deciding whether the subject should remain on investigational product and ultimately continued participation in the study. A patient does not necessarily have to have investigational product interrupted if they have self-injurious behavior that would be classified as nonsuicidal self-injurious behavior.

8.1.2. Permanent Discontinuation from Investigational Product

Investigational product should be permanently discontinued if the patient requests to discontinue investigational product.

Discontinuation of the investigational product for abnormal liver tests should be considered by the investigator when a patient meets 1 of the following conditions after consultation with the Lilly-designated medical monitor:

- ALT or AST >8x ULN
- ALT or AST >5x ULN for more than 2 weeks
- ALT or AST >3x ULN and total bilirubin level (TBL) >2x ULN or international normalized ratio (INR) >1.5

- ALT or AST >3x ULN with the appearance of fatigue, nausea, vomiting, right upper-quadrant pain or tenderness, fever, and/or rash
- ALP >3x ULN
- ALP >2.5x ULN and TBL >2x ULN
- ALP >2.5x ULN with the appearance of fatigue, nausea, vomiting, right quadrant pain or tenderness, fever, and/or rash

NOTE: Patients who are discontinued from investigational product due to a hepatic event or liver test abnormality should have additional hepatic safety data collected via the hepatic safety eCRF.

Investigational product should be permanently discontinued if any of the following laboratory abnormalities are observed:

- white blood cell count <1000 cells/µL ($1.00x10^3$/µL or 1.00 GI/L)
- ANC <500 cells/µL ($0.50x10^3$/µL or 0.50 GI/L)
- lymphocyte count <200 cells/µL ($0.20x10^3$/µL or 0.20 GI/L)
- hemoglobin <6.5 g/dL (<65.0 g/L)

NOTE: Temporary interruption rules (Section 8.1.1) must be followed where applicable. For laboratory values that meet permanent discontinuation thresholds, investigational product should be discontinued. However, if in the opinion of the investigator the laboratory abnormality is due to intercurrent illness such as cholelithiasis or another identified factor, laboratory tests may be repeated. Only when the laboratory value meets resumption thresholds (Table JAIY.4) following the resolution of the intercurrent illness or other identified factor, may the investigator restart investigational product, after consultation with the Lilly-designated medical monitor.

In addition, patients will be discontinued from investigational product in the following circumstances:

- pregnancy
- malignancy (except for successfully treated basal or squamous cell skin carcinoma)
- hepatitis B virus DNA is detected with a value above limit of quantitation or 2 sequential tests return a value of below the limit of quantitation (Section 9.4.8)
- develop a second VTE
- certain prohibited medications are taken per Section 7.7.3 (Prohibited Medications and Procedures)

If a patient discontinues investigational product for any reason, the patient is encouraged to remain in the study through Week 16 (Visit 8) and follow the regular visit schedule to provide the primary efficacy and safety data. Patients discontinuing from the investigational product prematurely for any reason should complete AE and other follow-up procedures per Section 2 (Schedule of Activities), Section 9.2 (Adverse Events), and Section 9.4 (Safety) of this protocol.

8.1.3. Discontinuation of Inadvertently Enrolled Patients

If the sponsor or investigator identifies a patient who did not meet enrollment criteria and was inadvertently enrolled, then the patient should be discontinued from study treatment unless there are extenuating circumstances that make it medically necessary for the patient to continue on study treatment. If the investigator and the sponsor clinical research physician agree it is medically appropriate to continue, the investigator must obtain documented approval from the sponsor clinical research physician to allow the inadvertently enrolled patient to continue in the study with or without treatment with investigational product. Safety follow-up is as outlined in Section 2 (Schedule of Activities), Section 9.2 (Adverse Events), and Section 9.4 (Safety) of this protocol.

8.2. Discontinuation from the Study

Patients may choose to withdraw from the study for any reason at any time, and the reason for early withdrawal will be documented.

Some possible reasons that may lead to permanent discontinuation include the following:

- enrollment in any other clinical study involving an investigational product or enrollment in any other type of medical research judged not to be scientifically or medically compatible with this study
- participation in the study needs to be stopped for medical, safety, regulatory, or other reasons consistent with applicable laws, regulations, and good clinical practice (GCP)
- investigator decision
  - The investigator decides that the patient should be discontinued from the study
  - If the patient, for any reason, requires treatment with another therapeutic agent (not allowed as part of rescue therapy [Section 7.7.4]) that has been demonstrated to be effective for treatment of the study indication, discontinuation from the study occurs prior to introduction of the new agent
- patient decision
  - The patient requests to be withdrawn from the study

Patients discontinuing from the study prematurely for any reason should complete AE and other safety follow-up per Section 2 (Schedule of Activities), Section 9.2 (Adverse Events), and Section 9.4 (Safety) of this protocol.

8.3. Lost to Follow-up

A patient will be considered lost to follow-up if he or she repeatedly fails to return for scheduled visits and is unable to be contacted by the study site. Site personnel are expected to make diligent attempts to contact patients who fail to return for a scheduled visit or were otherwise unable to be followed up by the site.

9. Study Assessments and Procedures

Section 2 lists the Schedule of Activities, with the study procedures and their timing (including tolerance limits for timing).

Appendix 2 and Appendix 4 list the laboratory tests that will be performed for this study.

Unless otherwise stated in the subsections below, all samples collected for specified laboratory tests will be destroyed within 60 days of receipt of confirmed test results. Certain samples may be retained for a longer period, if necessary, to comply with applicable laws, regulations, or laboratory certification standards.

9.1. Efficacy Assessments

9.1.1. Primary Efficacy Assessments

Validated Investigator's Global Assessment for Atopic Dermatitis (vIGA-AD): The IGA used in this study, the vIGA-AD (referred to as the IGA throughout the protocol) measures the investigator's global assessment of the patient's overall severity of their AD, based on a static, numeric 5-point scale from 0 (clear skin) to 4 (severe disease). The score is based on an overall assessment of the degree of erythema, papulation/induration, oozing/crusting, and lichenification.

9.1.2. Secondary Efficacy Assessments

9.1.2.1. Eczema Area and Severity Index Scores

The EASI assesses extent of disease at 4 body regions and measures 4 clinical signs: (1) erythema, (2) induration/papulation, (3) excoriation, and (4) lichenification each on a scale of 0 to 3. The EASI confers a maximum score of 72. The EASI evaluates 2 dimensions of AD: disease extent and clinical signs (Hanifin et al. 2001).

Body surface area affected by AD will be derived from data collected as part of the EASI assessment.

9.1.2.2. SCORing Atopic Dermatitis

The SCORAD index uses the rule of nines to assess disease extent and evaluates 6 clinical characteristics to determine disease severity: (1) erythema, (2) edema/papulation, (3) oozing/crusts, (4) excoriation, (5) lichenification, and (6) dryness. The SCORAD index also assesses subjective symptoms of pruritus and sleep loss. These 3 aspects: extent of disease, disease severity, and subjective symptoms combine to give a maximum possible score of 103 (Stalder et al. 1993; Kunz et al. 1997; Schram et al. 2012).

9.1.2.3. Hospital Anxiety Depression Scale

The Hospital Anxiety Depression Scale (HADS) is a 14-item self-assessment scale that determines the levels of anxiety and depression that a patient is experiencing over the past week. The HADS utilizes a 4-point Likert scale (e.g., 0 to 3) for each question and is intended for ages 12 to 65 years (Zigmond and Snaith 1983; White et al. 1999). Scores for each domain (anxiety

and depression) can range from 0 to 21, with higher scores indicating greater anxiety or depression (Zigmond and Snaith 1983; Snaith 2003).

9.1.3. Health Outcomes and Quality-of-Life Measures

The patient self-reported questionnaires will be administered via either an electronic patient diary or via an electronic tablet and in countries where the questionnaires have been translated into the native language of the region and linguistically validated.

9.1.3.1. Patient-Oriented Eczema Measure

The POEM is a simple, 7-item, patient-administered scale that assesses disease severity in children and adults. Patients respond to questions about the frequency of 7 symptoms (itching, sleep disturbance, bleeding, weeping/oozing, cracking, flaking, and dryness/roughness) over the last week. Response categories include "No days," "1-2 days," "3-4 days," "5-6 days," and "Every day" with corresponding scores of 0, 1, 2, 3, and 4, respectively. Scores range from 0 to 28 with higher total scores indicating greater disease severity (Charman et al. 2004).

9.1.3.2. Itch Numeric Rating Scale

The Itch Numeric Rating Scale (NRS) is a patient-administered, 11-point horizontal scale anchored at 0 and 10, with 0 representing "no itch" and 10 representing "worst itch imaginable." Overall severity of a patient's itching is indicated by selecting the number that best describes the worst level of itching in the past 24 hours (Naegeli et al. 2015; Kimball et al. 2016).

9.1.3.3. Atopic Dermatitis Sleep Scale

The Atopic Dermatitis Sleep Scale (ADSS) is a 3-item, patient-administered questionnaire developed to assess the impact of itch on sleep including difficulty falling asleep, frequency of waking, and difficulty getting back to sleep last night. Patient's rate their difficulty falling asleep and difficulty getting back to sleep, items 1 and 3, respectively, using a 5-point Likert-type scale with response options ranging from 0 "not at all" to 4 "very difficult." Patients report their frequency of waking last night, item 2, by selecting the number of times they woke up each night, ranging from 0 to 29 times. The ADSS is designed to be completed each day with respondents thinking about sleep "last night." Each item is scored individually.

9.1.3.4. Skin Pain Numeric Rating Scale

Skin Pain NRS is a patient-administered, 11-point horizontal scale anchored at 0 and 10, with 0 representing "no pain" and 10 representing "worst pain imaginable." Overall severity of a patient's skin pain is indicated by selecting the number that best describes the worst level of skin pain in the past 24 hours.

9.1.3.5. Patient Global Impression of Severity

The Patient Global Impression of Severity–Atopic Dermatitis (PGI-S-AD) is a single-item question asking the patient how they would rate their overall AD symptoms over the past 24 hours. The 5 categories of responses range from "no symptoms" to "severe."

9.1.3.6. Dermatology Life Quality Index

The Dermatology Life Quality Index (DLQI) is a simple, patient-administered, 10-item, validated, quality-of-life questionnaire that covers 6 domains including symptoms and feelings,

daily activities, leisure, work and school, personal relationships, and treatment. The recall period of this scale is over the "last week." Response categories include "not at all," "a lot," and "very much," with corresponding scores of 1, 2, and 3, respectively, and unanswered ("not relevant") responses scored as 0. Scores range from 0 to 30 with higher scores indicating greater impairment of quality of life. A DLQI total score of 0 to 1 is considered as having no effect on a patient's health-related QoL (Hongbo et al. 2005), and a 4-point change from baseline is considered as the minimal clinically important difference threshold (Khilji et al. 2002; Basra et al. 2015).

9.1.3.7. European Quality of Life–5 Dimensions–5 Levels

The European Quality of Life-5 Dimensions-5 Levels (EQ-5D-5L) is a standardized measure of health status that provides a simple, generic measure of health for clinical and economic appraisal. The EQ-5D-5L consists of 2 components: a descriptive system of the respondent's health and a rating of his or her current health state using a 0 to 100 mm Visual Analog Scale (VAS). The descriptive system comprises the following 5 dimensions: mobility, self-care, usual activities, pain/discomfort, and anxiety/depression. Each dimension has 5 levels: no problems, slight problems, moderate problems, severe problems, and extreme problems. The respondent is asked to indicate his or her health state by ticking (or placing a cross) in the box associated with the most appropriate statement in each of the 5 dimensions. It should be noted that the numerals 1 to 5 have no arithmetic properties and should not be used as an ordinal score. The VAS records the respondent's self-rated health on a vertical VAS where the endpoints are labeled "best imaginable health state" and "worst imaginable health state." This information can be used as a quantitative measure of health outcome. The EQ-5D-5L health states, defined by the EQ-5D-5L descriptive system, may be converted into a single summary index by applying a formula that essentially attaches values (also called weights) to each of the levels in each dimension (Herdman et al. 2011; EuroQol Group 2015 [WWW]).

9.1.3.8. Work Productivity and Activity Impairment Questionnaire–Atopic Dermatitis

The Work Productivity and Activity Impairment Questionnaire–Atopic Dermatitis (WPAI-AD) records impairment due to AD during the past 7 days. The WPAI-AD consists of 6 items grouped into 4 domains: absenteeism (work time missed), presenteeism (impairment at work/reduced on-the-job effectiveness), work productivity loss (overall work impairment/absenteeism plus presenteeism), and activity impairment. Scores are calculated as impairment percentages (Reilly et al. 1993), with higher scores indicating greater impairment and less productivity.

9.1.3.9. PROMIS Itch Questionnaire General Short Form 8a v1.0

Patient-Reported Outcomes Measurement Information System (PROMIS) is a set of person-centered measures that evaluates and monitors physical, mental, and social health in adults and children. It can be used with the general population and with individuals living with chronic conditions. The Itch General Short Form 8a (PIQ – General) within the PROMIS itch item bank consists of 8 items assessing the impact of itch on various aspects of life. Response

options range from 1=Never, 2=Rarely; 3=Sometimes; 4=Often; to 5=Almost always; total raw scores are converted to T-Scores with higher scores representing greater impact because of itch.

9.1.3.10. PROMIS Itch Questionnaire Activity and Clothing Short Form 8a v1.0

The Activity and Clothing Short Form 8a (PIQ – Activity and Clothing) within the PROMIS itch item bank consists of 8 items assessing activity and clothing related quality of life impairment from itch in adults "in the past 7 days". Response options range from 1=Never, 2=Rarely; 3=Sometimes; 4=Often; to 5=Almost always; total raw scores are converted to T-Scores with higher scores representing greater impact on activity and clothing because of itch.

9.1.3.11. PROMIS Itch Questionnaire Mood and Sleep Short Form 8a v1.0

The Mood and Sleep Short Form 8a (PIQ – Mood and Sleep) within the PROMIS Itch item bank consists of 8 items assessing mood and sleep related quality of life impairment from itch and impact of itch "in the past 7 days". Response options range from 1=Never, 2=Rarely; 3=Sometimes; 4=Often; to 5=Almost always; total raw scores are converted to T-Scores with higher scores representing greater impact on mood and sleep because of itch.

9.1.3.12. PROMIS Itch Questionnaire Scratching Behavior Short Form 5a v1.0

The Scratching Behavior Short Form 5a (PIQ – Scratching Behavior) within the PROMIS Itch item bank consists of 5 items assessing quality of life impairment from scratching behavior and the physical manifestations of itch in adults "in the past 7 days". Response options for the frequency of scratching behaviors range from 1=Never, 2=Rarely, 3=Sometimes; 4=Often, to 5=Almost always. The response options for the worry related to scratching items range from 1=Not at all; 2=A little bit; 3=Somewhat; 4=Quite a bit; to 5=Very much. Total raw scores are converted to T-Scores with higher scores representing more scratching behavior.

9.1.3.13. PROMIS Sleep Related Impairment Short Form 8a v1.0

The Sleep Related Impairment Short Form 8a (PROMIS – Sleep Impairment [PROMIS 2015]) within the PROMIS bank consists of 8 items measuring self-reported perceptions of alertness, sleepiness, and tiredness during usual waking hours, and the perceived functional impairments during wakefulness associated with sleep problems or impaired alertness "in the past 7 days". Response options range from 1=Not at all; 2=A little bit; 3=Somewhat; 4=Quite a bit; to 5=Very much. Total raw scores are converted to T-Scores with higher scores representing greater sleep impairment.

9.1.3.14. Neuro-QoL Cognitive Function Short Form v2.0

Neuro-QoL is a set of self-reported measures that assess the health-related quality of life (HRQOL) of adults and children with neurological disorders and is comprised of item banks and scales that evaluate symptoms, concerns, and issues that are relevant across disorders along with instruments that assess areas most relevant for specific patient populations (NINDS 2015). The Cognitive Function Short Form v2.0 (Neuro-QoL – Cognitive Function) domain within Neuro-QoL bank consists of 8 items measuring Executive Function (perceived difficulties in applications of mental health function related to planning, organizing, calculating, remembering and learning) "in the past 7 days" and General Concerns (perceived difficulties in everyday cognitive abilities such as memory, attention, and decision making) using the lead-in phrase

"how much difficulty do you currently have…". The response options for the Executive Function items range from 1=Very Often (several times a day); 2=Often (once a day); 3=Sometimes (2-3 times); 4=Rarely (once); to 5=Never). The response options for the General Concerns items range from 1=Cannot do; 2=A lot; 3=Somewhat; 4=A little, to 5=None. The total raw scores and are converted to T-Scores with higher scores indicating better (desirable) self-reported health.

9.1.3.15. Patient Benefit Index

The Patient Benefit Index (PBI) measures patient-defined treatment objectives and benefits, particularly in the course of a treatment (Augustin et al. 2009; Blome et al. 2011). It consists of 2 questionnaires. Before therapy, patients complete the standardized "Patient Needs Questionnaire" indicating individual importance of treatment objectives. This reflects their personal preferences with respect to therapeutic benefit. During the study, patients rate the extent to which the treatment objectives have been achieved in the "Patient Benefit Questionnaire". Response options range from 0=not at all, 1=somewhat; 2=moderately; 3=quite; 4=very; to 5=does/did not apply to me. A global score is calculated for each patient by weighing the achievement values of the treatment objectives by their importance to the individual patient. Moreover, the PBI will be supplemented by 6 rating scales assessing the following areas: physical well-being, emotional well-being, performance capacity on the job and in everyday living, social contacts, leisure activities, and quality of life. Patients with PBI $\geq$1 are considered having at least minimum patient-relevant treatment benefit.

9.1.4. Appropriateness of Assessments

All assessments utilized in this study are standard, widely used, and generally recognized as reliable, accurate, and relevant except ADSS and Skin Pain NRS, which are currently being developed and validated according to regulatory guidances.

9.2. Adverse Events

Investigators are responsible for monitoring the safety of patients who have entered this study and for alerting Lilly or its designee to any event that seems unusual, even if this event may be considered an unanticipated benefit to the patient.

The investigator is responsible for the appropriate medical care of patients during the study.

Investigators must document their review of each laboratory safety report.

The investigator remains responsible for following, through an appropriate health care option, AEs that are serious or otherwise medically important, considered related to the investigational product or the study, or that caused the patient to discontinue the investigational product before completing the study. The patient should be followed until the event resolves or stabilizes with appropriate diagnostic evaluation; for events that are not anticipated to resolve or stabilize, the patient should be followed until the treating physician (in consultation with the sponsor) determines that appropriate followup has been completed. The frequency of follow-up evaluations of the AE is left to the discretion of the investigator.

Lack of drug effect is not an AE in clinical studies, because the purpose of the clinical study is to establish treatment effect.

After the ICF is signed, study site personnel will record via eCRF the occurrence and nature of each patient's preexisting conditions. In addition, site personnel will record any change in the condition(s) and any new conditions as AEs.

Investigators should record the following via eCRF for each AE: time of onset, time of termination, severity, and their assessment of the potential relatedness of each AE to investigational product.

The investigator will interpret and document whether or not an AE has a reasonable possibility of being related to study treatment, study device, or a study procedure, taking into account the disease, concomitant treatment, or pathologies. A "reasonable possibility" means that there is a cause-and-effect relationship between the investigational product, study device, and/or study procedure and the AE. The investigator answers yes/no when making this assessment.

Planned surgeries and nonsurgical interventions should not be reported as AEs unless the underlying medical condition has worsened during the course of the study.

If a patient's investigational product is discontinued as a result of an AE, study site personnel must report this to Lilly or its designee via eCRF, clarifying if possible the circumstances leading to any dosage modifications, or discontinuations of treatment.

9.2.1. Serious Adverse Events

An SAE is any AE from this study that results in 1 of the following outcomes:

- death
- initial or prolonged inpatient hospitalization
- a life-threatening experience (i.e., immediate risk of dying)
- persistent or significant disability/incapacity
- congenital anomaly/birth defect
- important medical events that may not be immediately life threatening or result in death or hospitalization but may jeopardize the patient or may require intervention to prevent 1 of the other outcomes listed in the definition above. Examples of such medical events include allergic bronchospasm requiring intensive treatment in an emergency room or at home, blood dyscrasias or convulsions that do not result in inpatient hospitalization, or the development of drug dependency or drug abuse.

All AEs occurring after signing the ICF are recorded in the eCRF and assessed for serious criteria. The SAE reporting to the sponsor begins after the patient has signed the ICF and has received investigational product. However, if an SAE occurs after signing the ICF, but prior to receiving investigational product, the SAE should be reported to the sponsor as per SAE reporting requirements and timelines if it is considered reasonably possibly related to study procedure.

Study site personnel must alert Lilly or its designee of any SAE within 24 hours of investigator awareness of the event via a sponsor-approved method (e.g. investigator space portal, telephone, or fax). If alerts are issued via telephone, they are to be immediately followed with official notification on study-specific SAE forms. This 24-hour notification requirement refers to the initial SAE information and all follow-up SAE information. Patients with a serious hepatic AE should have additional data collected using the hepatic safety eCRF. Investigators can contact the sponsor via telephone at any time using the qualified medical personnel or Lilly affiliate medical contact details which are provided in the site study file.

Pregnancy (during maternal or paternal exposure to investigational product) does not meet the definition of an AE. However, to fulfill regulatory requirements, any pregnancy should be reported following the SAE process to collect data on the outcome for both mother and fetus.

Investigators are not obligated to actively seek AEs or SAEs in subjects once they have discontinued and/or completed the study (the patient summary CRF has been completed). However, if the investigator learns of any SAE, including a death, at any time after a subject has been discharged from the study, and he or she considers the event reasonably possibly related to the study treatment or study participation, the investigator must promptly notify Lilly.

9.2.1.1. Suspected Unexpected Serious Adverse Reactions

Suspected unexpected serious adverse reactions (SUSARs) are serious events that are not listed in the IB and that the investigator identifies as related to investigational product or procedure. United States 21 CFR 312.32 and European Union Clinical Trial Directive 2001/20/EC and the associated detailed guidances or national regulatory requirements in participating countries require the reporting of SUSARs. Lilly has procedures that will be followed for the recording and expedited reporting of SUSARs that are consistent with global regulations and the associated detailed guidances.

9.2.2. Adverse Events of Special Interest

Adverse events of special interest will include the following:

- infections (including TB, herpes zoster, or opportunistic infections)
- malignancies (except for successfully treated basal or squamous cell carcinoma)
- hepatic events (Section 9.4.9)
- major adverse cardiovascular events (MACE) (Section 9.4.10)
- thrombotic events (such as deep vein thrombosis [DVT] and pulmonary embolism)

Sites will provide details on these AEs as instructed on the eCRF and may be asked for additional description by Lilly.

9.2.3. Complaint Handling

Lilly collects product complaints on investigational products and drug delivery systems used in clinical studies to ensure the safety of study participants, monitor quality, and to facilitate process and product improvements.

Patients will be instructed to contact the investigator as soon as possible if he or she has a complaint or problem with the investigational product so that the situation can be assessed.

9.3. Treatment of Overdose

Refer to the IB.

9.4. Safety

Any clinically significant findings from ECG testing, physical examination, vital signs measurements, or laboratory measurements that result in a diagnosis and that occur after the patient receives the first dose of study treatment should be reported to Lilly or its designee as an AE via eCRF.

9.4.1. Electrocardiograms

A single 12-lead standard ECG will be obtained locally at Visit 1 and read by a qualified physician (the investigator or qualified designee) at the site to determine whether the patient meets entry criteria.

Electrocardiograms may be obtained at additional times, when deemed clinically necessary.

9.4.2. Vital Signs

For each patient, vital signs should be measured according to the Schedule of Activities (Section 2).

9.4.3. Physical Examination

For each patient, a complete physical examination (excluding pelvic and rectal examinations) will be performed at Visit 1 (Screening). A symptom-directed physical examination will be performed at other visits as specified in the Schedule of Activities (Section 2). A complete physical examination may be repeated at the investigator's discretion at any time a patient presents with physical complaints.

9.4.4. Laboratory Tests

For each patient, laboratory tests detailed in Appendix 2 should be conducted according to the Schedule of Activities (Section 2). With the exception of laboratory test results that may unblind the study, Lilly or its designee will provide the investigator with the results of laboratory tests analyzed by a central vendor, if a central vendor is used for the clinical trial.

9.4.5. Columbia Suicide Severity Rating Scale

The C-SSRS captures the occurrence, severity, and frequency of suicidal ideation and/or behavior during the assessment period. The scale includes suggested questions to solicit the type of information needed to determine if suicidal ideation and/or behavior occurred. The C-SSRS is administered by an appropriately trained health care professional with at least 1 year of patient care/clinical experience. The tool was developed by the National Institute of Mental Health trial group for the purpose of being a counterpart to the Columbia Classification Algorithm of Suicide

Assessment categorization of suicidal events. For this study, the scale has been adapted (with permission from the scale authors) to include only the portion of the scale that captures the occurrence of the 11 preferred ideation and behavior categories.

The nonleading AE collection should occur prior to the collection of the C-SSRS. If a suicide-related event is discovered *during the C-SSRS*, but was not captured during the nonleading AE collection, sites should not change the AE form. If an event is serious or leads to discontinuation, this is an exception where the SAE and/or AE leading to discontinuation should be included on the AE form and the process for reporting SAEs should be followed.

9.4.6. Self-Harm and Follow-up Supplement Forms

Suicide-related events (behavior and/or ideations) will be assessed and evaluated at every visit with the administration of the C-SSRS and the Self-Harm Supplement Form. The Self-Harm Supplement Form is a single question to enter the number of suicidal behavior events, possible suicide behaviors, or nonsuicidal self-injurious behaviors. If the number of behavioral events is greater than 0, it will lead to the completion of the self-harm follow-up form. The self-harm follow-up form is a series of questions that provides a more detailed description of the behavior cases.

9.4.7. Chest x-ray and Tuberculosis Testing

A posterior–anterior view chest x-ray will be obtained locally at screening (Visit 1), unless results from a chest x-ray obtained within 6 months prior to the study are available. The chest x-ray will be reviewed by the investigator or his or her designee to exclude patients with active TB infection. In addition, patients will be tested at screening (Visit 1) for evidence of active or latent TB as described in the exclusion criteria, Section 6.2.

Investigators should follow local guidelines for monitoring patients for TB if a patient is at high risk for acquiring TB or reactivation of latent TB.

9.4.8. Hepatitis B Virus DNA Monitoring

Patients who are HBcAb positive and HBV DNA negative (undetectable) at Visit 1 will require measurement of HBV DNA at Week 16 (Visit 8) or early termination visit, regardless of their hepatitis B surface antibody (HBsAb) status.

The following actions should be taken in response to HBV DNA test results:

- If a single result is obtained with a value “below limit of quantitation,” the test should be repeated within approximately 2 weeks. If the repeat test result is “target not detected,” monitoring will resume for those patients enrolling in the long-term extension study, JAHN.
- If the patient has 2 or more test results with a value “below limit of quantitation” or a test result above the limit of quantitation, the patient will be permanently discontinued from investigational product (Section 8.1.2) and should be referred to a hepatology specialist.

9.4.9. Hepatic Safety Monitoring and Data Collection

If a study patient experiences elevated ALT ≥3x ULN, ALP ≥2x ULN, or elevated TBL ≥2x ULN, liver testing (Appendix 4) should be repeated within 3 to 5 days including ALT, AST, ALP, TBL, direct bilirubin, gamma-glutamyl transferase, and creatine kinase to confirm the abnormality and to determine if it is increasing or decreasing. If the abnormality persists or worsens, clinical and laboratory monitoring should be initiated by the investigator and in consultation with the study medical monitor. Monitoring of ALT, AST, TBL, and ALP should continue until levels normalize or return to approximate baseline levels.

Discontinuation criteria of investigational products, either temporary interruption or permanent discontinuation, due to abnormal ALT, AST, TBL, or ALP, are detailed in Section 8.1.

Additional safety data should be collected via the hepatic eCRF if 1 or more of the following conditions occur:

- elevation of serum ALT to ≥5x ULN on 2 or more consecutive blood tests
- elevated serum TBL to ≥ 2x ULN (except for cases of known Gilbert's syndrome)
- elevation of serum ALP to ≥2x ULN on 2 or more consecutive blood tests
- patient discontinued from treatment due to a hepatic event or abnormality of liver tests
- hepatic event considered to be an SAE

See Appendix 4 and Appendix 5 for a description of hepatic laboratory values that warrant exclusion from the study, temporary or permanent discontinuation of investigational product, or additional safety collection via the hepatic eCRF.

9.4.10. Safety Monitoring

Lilly will periodically review evolving aggregate safety data within the study by appropriate methods.

In the event that safety monitoring uncovers an issue that needs to be addressed by unblinding at the group level, only members of the DMC (an advisory group for this study formed to protect the integrity of data [refer to Interim Analyses section, Section 10.3.7]) can conduct additional analyses of the safety data.

The Lilly clinical research physician will monitor safety data throughout the course of the study. Lilly will review SAEs within time frames mandated by company procedures. The Lilly clinical research physician will, as is appropriate, consult with the functionally independent Global Patient Safety therapeutic area physician or clinical scientist and periodically review trends in safety data and laboratory analytes. Any concerning trends in frequency or severity noted by an investigator and/or Lilly (or designee) may require further evaluation.

All deaths and SAE reports will be reviewed in a blinded manner by Lilly during the clinical trial. These reports will be reviewed to ensure completeness and accuracy, but will not be unblinded to Lilly during the clinical trial. If a death or a clinical AE is deemed serious, unexpected, and possibly related to investigational product, only Lilly Global Patient Safety will be unblinded for regulatory reporting and safety monitoring purposes. These measures will

preserve the integrity of the data collected during this trial and minimize any potential for bias while providing for appropriate safety monitoring.

Investigators will monitor vital signs and carefully review findings that may be associated with cardiovascular events and VTE (Appendix 6). Adverse event reports and vital signs will be collected at each study visit. The cardiovascular monitoring plan includes the following:

- regular monitoring of lipid levels
- potential MACE (cardiovascular death, MI, and stroke), other cardiovascular events (such as hospitalization for unstable angina, hospitalization for heart failure, serious arrhythmia, resuscitated sudden death, cardiogenic shock, coronary revascularization such as coronary artery bypass graft or percutaneous coronary intervention), venous thrombotic events and noncardiovascular deaths will be identified by the investigative site or through medical review and will be sent to a blinded Clinical Event Committee for adjudication at regular intervals.

9.5. Pharmacokinetics

Not applicable.

9.6. Pharmacodynamics

Not applicable.

9.7. Pharmacogenetics

9.7.1. Blood Samples for Pharmacogenetic Research

A blood sample will be collected for pharmacogenetic analysis as specified in the Schedule of Activities (Section 2) where local regulations allow.

There is growing evidence that genetic variation may impact a patient's response to therapy. Variable response to therapy may be due to genetic determinants that impact drug absorption, distribution, metabolism, and excretion, the mechanism of action of the drug, the disease etiology, and/or the molecular subtype of the disease being treated. In the event of an unexpected AE, the samples may be genotyped and analysis may be performed to evaluate a genetic association with response to baricitinib. These investigations may be limited to targeted exome sequencing approach of known targets involved in drug metabolism or, if appropriate, genome-wide association studies may be performed to identify regions of the genome associated with the variability observed in drug response. Samples will only be used for investigations related to disease and drug or class of drugs under study in the context of this clinical program.

Samples will not be used to conduct unspecified disease or population genetic research either now or in the future. Samples will be used to investigate variable response to baricitinib and to investigate genetic variants thought to play a role in AD or other inflammatory skin diseases. Assessment of variable response may include evaluation of AEs or differences in efficacy.

All samples will be coded with the patient number. These samples and any data generated can be linked back to the patient only by the investigator site personnel.

Samples will be retained for a maximum of 15 years after the last patient visit for the study, or for a shorter period if local regulations and/or (ethical review boards (ERBs)/investigational review boards) impose shorter time limits, at a facility selected by Lilly. This retention period enables use of new technologies, response to regulatory questions, and investigation of variable response that may not be observed until later in the development of baricitinib or after baricitinib becomes commercially available.

Molecular technologies are expected to improve during the 15-year storage period and therefore cannot be specifically named. However, existing approaches include whole genome or exome sequencing, genome-wide association studies, and candidate gene studies. Regardless of technology utilized, genotyping data generated will be used only for the specific research scope described in this section.

9.8. Biomarkers

Biomarker research is performed to address questions of relevance to drug disposition, target engagement, pharmacodynamics (PD), mechanism of action, variability of patient response (including safety), and clinical outcome. Sample collection is incorporated into clinical studies to enable examination of these questions through measurement of biomolecules including DNA, RNA, proteins, lipids, and other cellular elements.

Blood samples for nonpharmacogenetic biomarker research will be collected at the times specified in the Schedule of Activities (Section 2) where local regulations allow.

Samples will be used for research on the drug target, disease process, variable response to baricitinib, pathways associated with AD, mechanism of action of baricitinib, and/or research method or in validating diagnostic tools or assay(s) related to AD.

All samples will be coded with the patient number. These samples and any data generated can be linked back to the patient only by the investigator site personnel.

Samples will be retained for a maximum of 15 years after the last patient visit for the study, or for a shorter period if local regulations and ERBs impose shorter time limits, at a facility selected by Lilly. This retention period enables use of new technologies, response to regulatory questions, and investigation of variable response that may not be observed until later in the development of baricitinib or after baricitinib becomes commercially available.

9.9. Medical Resource Utilization and Health Economics

Health Economics will be evaluated in this study utilizing the EQ-5D-5L and WPAI-AD (Section 9.1.3). Medical Resource Utilization parameters will not be evaluated in this study.

10. Statistical Considerations

10.1. Sample Size Determination

Study JAIY will aim to enroll approximately 300 patients ≥18 years of age. The proposed sample size will ensure an 89% power to detect an absolute difference of 20% between the baricitinib 4-mg and placebo treatment groups and the baricitinib 2-mg and placebo treatment groups, each using a 2-sided alpha of 0.025 and a Fisher exact test, assuming a 10% placebo response rate for the primary endpoint. The assumptions are based on what was observed in the Phase 2 study (JAHG). The proposed end point of IGA 0 or 1 represents patients whose AD is clear or almost clear from a baseline of moderate or severe disease. The anticipated effect size represents 3 times more patients achieving this benefit compared to placebo, which, in discussion with therapeutic experts, is of a magnitude that is considered clinically relevant.

Sample size and power estimates were obtained from nQuery® Advisor 7.0.

10.2. Populations for Analyses

Unless otherwise specified, the efficacy and health outcome analyses will be conducted on the intent-to-treat population, defined as all randomized patients, even if the patient does not receive the correct treatment, or otherwise did not follow the protocol. Patients will be analyzed according to the treatment to which they were assigned. Significant protocol violations will be described in the SAP.

Safety analyses will be done on all randomized patients who receive at least 1 dose of investigational product and who did not discontinue from the study for the reason "Lost to Follow-up" at the first postbaseline visit.

Further details of other populations will be described in the SAP. Patients will be analyzed according to the dosing regimen to which they were assigned in the Treatment Period.

10.3. Statistical Analyses

10.3.1. General Statistical Considerations

Statistical analysis of this study will be the responsibility of Lilly or its designee. A detailed SAP describing the statistical methodologies will be developed by Lilly or its designee.

Any change to the data analysis methods described in the protocol will require an amendment ONLY if it changes a principal feature of the protocol. Any other change to the data analysis methods described in the protocol, and the justification for making the change, will be described in the clinical study report (CSR). Additional exploratory analyses of the data will be conducted as deemed appropriate.

All tests of treatment effects will be conducted at a 2-sided alpha level of 0.05, unless otherwise stated. Treatment comparisons of discrete efficacy variables between baricitinib and placebo will be made using a logistic regression analysis with region, disease severity, and treatment group in the model. The percentages, difference in percentages, and 95% confidence interval

(CI) of the difference in percentages will be reported. Treatment-by-region interaction will be added to the logistic regression model of the primary and key secondary variables as a sensitivity analysis. If this interaction is significant at a 2-sided 0.1 level, further inspection will be used to assess whether the interaction is quantitative (i.e., the treatment effect is consistent in direction but not size of effect) or qualitative (the treatment is beneficial for some but not all regions). The p-value from the Fisher exact test will also be produced.

When evaluating continuous measures over time, a restricted maximum likelihood-based mixed-effects model of repeated measures (MMRM) will be used. The model will include treatment, region, baseline severity, visit, and treatment-by-visit interaction as fixed categorical effects and baseline score and baseline score-by-visit interaction as fixed continuous effects. An unstructured (co)variance structure will be used to model the between- and within-patient errors. If this analysis fails to converge, other structures will be tested. The Kenward–Roger method will be used to estimate the degrees of freedom. Type III sums of squares for the least squares means (LSMs) will be used for the statistical comparison; 95% CI will also be reported. Contrasts will be set up within the model to test treatment groups at specific time points of interest. Further details on the use of MMRM will be described in the SAP.

Treatment comparisons of continuous efficacy and health outcome variables may also be made using analysis of covariance (ANCOVA) with region, disease severity, treatment group, and baseline value in the model. Type III tests for LSM will be used for statistical comparison between treatment groups. The LSM difference, standard error, p-value, and 95% CI may also be reported. The method used to handle missing data will be specified in the SAP.

Fisher exact test will be used for the AEs, discontinuation, and other categorical safety data for between-treatment group comparisons. Continuous vital signs, body weight, and other continuous safety variables including laboratory variables will be analyzed by an ANCOVA with treatment and baseline value in the model. Shift tables for categorical safety analyses (e.g., "high" or "low" laboratory results) will also be produced.

Missing data imputation:

1. Nonresponder imputation (NRI): All patients who discontinue the study or the study treatment at any time for any reason will be defined as nonresponders for the NRI analysis for categorical variables such as IGA 0/1 or EASI 50/75/90 after discontinuation and onward. Patients who receive rescue therapy will be analyzed as nonresponders after rescue and onward. An additional analysis will be performed that includes all available data whether rescue medication was given or not.
2. MMRM: Continuous variables such as EASI and SCORAD scores will be assumed to be missing after rescue or discontinuation and then an MMRM analysis will be performed. An additional analysis will be performed that includes all available data whether rescue medication was given or not.
3. Last observed carried forward (LOCF): An additional analysis will be performed that uses the last observed value on or prior to discontinuation or rescue therapy. This will

then be analyzed using a logistic model for categorical variables or ANCOVA for continuous variables as described above.

Additional sensitivity analyses for the primary and key secondary endpoints such as tipping point analyses as well as a reference based multiple imputation method may be done and will be specified in the SAP.

Adjustment for Multiple Comparisons:

Multiplicity controlled analyses will be performed on the primary and major secondary endpoints to control the overall family-wise Type I error rate at a 2-sided α level of 0.05. The graphical multiple testing procedure described in Bretz et al. (2011) will be used. The graphical approach is a closed testing procedure; hence it strongly controls the family-wise error rate across all endpoints (Alosh et al. 2014). Details of the specific graphical testing scheme (including testing order, interrelationships, Type I error allocation, and the associated propagation) will be prespecified in the SAP.

The following is a list of primary and key secondary endpoints to be tested:

Primary:

- proportion of baricitinib 4-mg patients achieving IGA of 0 or 1 and ≥2-point improvement from baseline at Week 16.
- proportion of baricitinib 2-mg patients achieving IGA of 0 or 1 and ≥2-point improvement from baseline at Week 16.

Key Secondaries:

Evaluated for 2-mg and 4-mg:

- proportion of patients achieving EASI75 at 16 weeks
- proportion of patients achieving EASI90 at 16 weeks
- percent change from baseline in EASI score at 16 weeks
- proportion of patients achieving SCORAD75 at 16 weeks
- proportions of patients achieving a 4-point improvement in Itch NRS at 2 days, 1 week, 2 weeks, 4 weeks, and 16 weeks
- mean change from baseline in the score of Item 2 from the ADSS at 1 week and 16 weeks
- mean change from baseline in skin pain NRS at 16 weeks

10.3.2. Treatment Group Comparability

10.3.2.1. Patient Disposition

All patients who discontinue from the study or the study treatment will be identified, along with their reason for discontinuation. Reasons for discontinuation from the study will be summarized by treatment group.

10.3.2.2. Patient Characteristics

Demographic and baseline characteristics will be summarized descriptively by treatment group. Descriptive statistics including number of patients, mean, standard deviation, median, minimum,

and maximum will be provided for continuous measures, and frequency counts and percentages will be tabulated for categorical measures. No formal statistical comparisons will be made among treatment groups unless otherwise stated.

10.3.2.3. Concomitant Therapy

Concomitant medications will be descriptively summarized by treatment group in terms of frequencies and percentages using the safety population. The medications will be coded accordingly.

10.3.2.4. Treatment Compliance

Treatment compliance with the randomly assigned study medication will be evaluated at every clinic visit through the counts of returned investigational product tablets. A patient will be considered significantly noncompliant if he or she misses more than 20% of the prescribed doses during the study, unless the patient's investigational product is withheld by the investigator for safety reasons; that is, compliance <80%. Similarly, a patient will be considered significantly noncompliant if he or she is judged by the investigator to have intentionally or repeatedly taken more than the prescribed amount of medication, that is, compliance ≥120%.

10.3.3. Efficacy Analyses

10.3.3.1. Primary Analyses

The primary efficacy measure is the binary outcome of response defined as IGA score of 0 or 1 (clear or almost clear skin) and ≥2-point improvement from baseline at Week 16. Primary analysis will be conducted using a logistic regression as described above with treatment and the stratification variables (disease severity and region) in the model. Nonresponder imputation for missing data as described above will be used.

Additional analysis of the primary efficacy outcome will include analyzing the outcome as observed, that is, whether or not rescue medication was used.

10.3.3.2. Secondary Analyses

The following secondary categorical outcomes will be analyzed in a similar manner as the primary; that is, using the same logistic regression model, but with respective baseline scores as further covariate included. Nonresponder imputation will be used for these analyses unless otherwise noted.

- EASI75 at Week 16. EASI75 is defined as having an improvement of at least 75% from baseline. Besides NRI, this outcome will also be analyzed using observed cases, that is, whether rescue medication was given or not.
- EASI90 at Week 16. EASI90 is defined as having an improvement of at least 90% from baseline.
- SCORAD75 at Week 16. SCORAD75 is defined as having an improvement of at least 75% from baseline.
- SCORAD90 at Week 16. SCORAD90 is defined as having an improvement of at least 90% from baseline.
- 4-point improvement in Itch NRS at 1 week, 2 weeks, 4 weeks, and 16 weeks.

The following continuous measures will be analyzed with the MMRM model described above unless otherwise noted. Contrasts within the MMRM model will be used to assess treatment differences for time points of interest as specified above in the list of objectives.

- Mean change from baseline in the following outcome measures:
    - ADSS Item 2 score
    - EASI score
    - SCORAD score
    - BSA
    - Itch NRS
    - POEM total score
    - PGI-S-AD
    - HADS
    - DLQI total score
    - WPAI (4 domains)
    - EQ-5D-5L (VAS, health state index)
    - PIQ – General
    - PIQ – Activity and Clothing
    - PIQ – Mood and Sleep
    - PIQ – Scratching Behavior
    - PROMIS – Sleep-Related Impairment
    - Neuro-QoL - Cognitive Function
    - Patient Benefit Index.

The EASI total score and SCORAD total score will also be analyzed as observed, that is, not assuming missing values after rescue medication is given.

10.3.4. Safety Analyses

All safety data will be descriptively summarized by treatment groups and analyzed using the safety population.

Treatment-emergent adverse events are defined as AEs that first occurred or worsened in severity after the first dose of study treatment. The number of TEAEs as well as the number and percentage of patients who experienced at least 1 TEAE will be summarized using MedDRA (Medical Dictionary for Regulatory Activities) for each system organ class (or a body system) and each preferred term by treatment group. Serious adverse events and AEs that lead to discontinuation of investigational product will also be summarized by treatment group. Fisher exact test will be used to perform comparisons between each baricitinib dose and the placebo group.

All clinical laboratory results will be descriptively summarized by treatment group. Individual results that are outside of normal reference ranges will be flagged in data listings. Quantitative clinical hematology, chemistry, and urinalysis variables obtained at the baseline to postbaseline visits will be summarized as changes from baseline by treatment group and analyzed using ANCOVA with treatment and baseline value in the model. Categorical variables, including the incidence of abnormal values and incidence of adverse events of special interest, will be

summarized by frequency and percentage of patients in corresponding categories. Shift tables will be presented for selected measures.

Observed values and changes from baseline (predose or screening if missing) for vital signs and physical characteristics will be descriptively summarized by treatment group and time point. Change from baseline to postbaseline in vital signs and body weight will be analyzed using ANCOVA with treatment and baseline value in the model.

The incidence and average duration of investigational product interruptions will be summarized and compared descriptively among treatment groups. Various techniques may be used to estimate the effects of investigational product interruptions on safety measures. Further analyses may be performed and will be planned in the SAP.

Data collected after initiation of rescue therapy will be summarized as appropriate.

10.3.5. Pharmacokinetic/Pharmacodynamic Analyses

Not applicable.

10.3.6. Other Analyses

10.3.6.1. Health Outcome Measures

The health outcome measures will be analyzed using methods described for continuous or categorical data as described for efficacy measures in Section 10.3.3.

10.3.6.2. Subgroup Analyses

To assess whether the treatment effect is similar across subgroups for the primary efficacy outcome, a logistic model will be used and will include treatment, stratification variables, the subgroup variable (e.g., sex) and the subgroup by treatment interaction. If the interaction is statistically significant at α=0.10, the nature of the interaction will be explored, that is, within each subgroup the treatment effect will be estimated. Similarly, for the continuous variables of EASI, the MMRM model will include additional variables for subgroup and the subgroup by treatment interaction.

Subgroups to be evaluated will include region, baseline severity, sex, age, race, prior therapy, etc. Further definitions for the levels of the subgroup variables, the analysis methodology, and any additional subgroup analyses will be defined in the SAP. As this study is not powered for subgroup analyses, all subgroup analyses will be treated as exploratory.

10.3.7. Interim Analyses

10.3.7.1. Data Monitoring Committee

A DMC will oversee the conduct of this trial. The DMC will consist of members external to Lilly. This DMC will follow the rules defined in the DMC charter, focusing on potential and identified risks for this molecule and for this class of compounds. Data Monitoring Committee membership will include, at a minimum, specialists with expertise in dermatology, statistics, and other appropriate specialties.

The DMC will be authorized to review unblinded results of analyses by treatment group prior to database lock, including study discontinuation data, AEs including SAEs, clinical laboratory data, vital sign data, etc. The DMC may recommend continuation of the study, as designed; temporary suspension of enrollment; or the discontinuation of a particular dose regimen or the entire study. While the DMC may request to review efficacy data to investigate the benefit/risk relationship in the context of safety observations for ongoing patients in the study, no information regarding efficacy will be communicated. Moreover, the study will not be stopped for positive efficacy results nor will it be stopped for futility. Hence, no alpha is spent. Details of the DMC, including its operating characteristics, will be documented in a DMC charter and DMC analysis plan.

Besides DMC members, a limited number of pre-identified individuals may gain access to the limited unblinded data, as specified in the unblinding plan, prior to the final database lock for preparation of regulatory documents. Information that may unblind the study during the analyses will not be reported to study sites or blinded study team until the study has been unblinded.

Unblinding details will be specified in a separate unblinding plan document.

10.3.7.2. Adjudication Committee

A blinded Clinical Event Committee will adjudicate potential MACE (cardiovascular death, MI, and stroke), other cardiovascular events (such as hospitalization for unstable angina, hospitalization for heart failure, serious arrhythmia, resuscitated sudden death, cardiogenic shock, coronary revascularization such as coronary artery bypass graft or percutaneous coronary intervention), venous thrombotic events, and noncardiovascular deaths. Details of membership, operations, recommendations from the Committee, and the communication plan will be documented in the Charter.

11. References

Alosh M, Bretz F, Huque M. Advanced multiplicity adjustment methods in clinical trials. *Stat Med.* 2014;33(4):693-713.

Augustin M, Radtke MA, Zschocke I, Blome C, Behechtnejad J, Schäfer I, Reusch M, Mielke V, Rustenbach SJ. The patient benefit index: a novel approach in patient-defined outcomes measurement for skin diseases. *Arch Dermatol Res*. 2009;301(8):561-571.

Basra MK, Salek MS, Camilleri L, Sturkey R, Finlay AY. Determining the minimal clinically important difference and responsiveness of the Dermatology Life Quality Index (DLQI): further data. *Dermatology*. 2015;230(1):27-33.

Bieber, T. Atopic dermatitis. *Ann Derm*. 2010;22(2):125-137.

Bieber T, Straeter B. Off-label prescriptions for atopic dermatitis in Europe. *Allergy*. 2015;70(1):6-11.

Blome C, Augustin M, Behechtnejad J, Rustenbach SJ. Dimensions of patient needs in dermatology: subscales of the patient benefit index. *Arch Dermatol Res.* 2011;303(1):11-17.

Bretz F, Posch M, Glimm E, Klinglmueller F, Maurer W, Rohmeyer K. Graphical approaches for multiple comparison procedures using weighted Bonferroni, Simes, or parametric tests. *Biom J*. 2011;53(6):894-913.

Charman CR, Venn AJ, Williams, HC. The patient-oriented eczema measure. *Arch Dermatol.* 2004;140:1513-1519.

Clark JD, Flanagan ME, Telliez JB. Discovery and development of Janus kinase (JAK) inhibitors for inflammatory diseases. *J Med Chem*. 2014;57(12):5023-5038.

Divekar R, Kita H. Recent advances in epithelium-derived cytokines (IL-33, IL-25, and thymic stromal lymphopoietin) and allergic inflammation. *Curr Opin Allergy Clin Immunol*. 2015;15(1):98-103.

Dougados M, van der Heijde D, Chen YC, Greenwald M, Drescher E, Liu J, Beattie S, Witt S, de la Torre I, Gaich C, Rooney T, Schlichting D, de Bono S, Emery P. Baricitinib in patients with inadequate response or intolerance to conventional synthetic DMARDs: results from the RA-BUILD study. *Ann Rheum Dis*. 2017;76(1):88-95.

Eichenfield LF, Tom WL, Chamlin SL, Feldman SR, Hanifin JM, Simpson EL, Berger TG, Bergman JN, Cohen DE, Cooper KD, Cordoro KM, Davis DM, Krol A, Margolis DJ, Paller AS, Schwarzenberger K, Silverman RA, Williams HC, Elmets CA, Block J, Harrod CG, Smith Begolka W, Sidbury R. Guidelines of care for the management of atopic dermatitis: section 1. Diagnosis and assessment of atopic dermatitis. *J Am Acad Dermatol*. 2014;70(2):338-351.

EuroQol Group. EQ-5D-5L User Guide. Version 2.1. Available at: http://www.euroqol.org/fileadmin/user_upload/Documenten/PDF/Folders_Flyers/EQ-5D-5L_UserGuide_2015.pdf. April 2015. Accessed March 2017.

Fleischmann R, Schiff M, van der Heijde D, Ramos-Remus C, Spindler A, Stanislav M, Zerbini CA, Gurbuz S, Dickson C, de Bono S, Schlichting D, Beattie S, Kuo WL, Rooney T, Macias W, Takeuchi T. Baricitinib, methotrexate, or combination in patients with rheumatoid arthritis

and no or limited prior disease-modifying antirheumatic drug treatment. *Arthritis Rheumatol*. 2017;69(3):506-517.

Fridman JS, Scherle PA, Collins R, Burn TC, Li Y, Li J, Covington MB, Thomas B, Collier P, Favata MF, Wen X, Shi J, McGee R, Haley PJ, Shepard S, Rodgers JD, Yeleswaram S, Hollis G, Newton RC, Metcalf B, Friedman SM, Vaddi K. Selective inhibition of JAK1 and JAK2 is efficacious in rodent models of arthritis: preclinical characterization of INCB028050. *J Immunol*. 2010;184(9):5298-5307.

Futamura M, Leshem YA, Thomas KS, Nankervis H, Williams HC, Simpson EL. A systematic review of Investigator Global Assessment (IGA) in atopic dermatitis (AD) trials: Many options, no standards. *J Am Acad Dermatol*. 2016;74(2):288-294.

Genovese MC, Kremer J, Zamani O, Ludivico C, Krogulec M, Xie L, Beattie SD, Koch AE, Cardillo TE, Rooney TP, Macias WL, de Bono S, Schlichting DE, Smolen JS. Baricitinib in patients with refractory rheumatoid arthritis. *N Engl J Med*. 2016;374(13):1243-1252.

Guttman-Yassky E, Silverberg JI, Nemoto O, Forman SB, Wilke A, Prescilla R, de la Peña A, Nunes FP, Janes J, Gamalo M, Donley D, Paik J, DeLozier AM, Nickoloff BJ, Simpson EL. Baricitinib in adult patients with moderate to severe atopic dermatitis: a phase 2 parallel, double-blinded, randomized placebo-controlled multiple-dose study. *J Am Acad Derm*. 2018;pii: S0190-9622(18)30129-4.

Hanifin JM, Thurston M, Omoto M, Cherill R, Tofte SJ, Graeber M. The eczema area and severity index (EASI): assessment of reliability in atopic dermatitis. EASI Evaluator Group. *Exp Dermatol*. 2001;10(1):11-18.

Herdman M, Gudex C, Lloyd A, Janssen MF, Kind P, Parkin D, Bonsel G, Badia X. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). *Qual Life Res*. 2011;20(10):1727-1736.

Hongbo Y, Thomas CL, Harrison MA, Salek MS, Finlay AY. Translating the science of quality of life into practice: What do dermatology life quality index scores mean? *J Invest Dermatol*. 2005;125(4):659-664.

Khilji FA, Gonzalez M, Finlay AY. Clinical meaning of change in Dermatology Life Quality Index scores. *Br J Dermatol*. 2002;147(suppl 62):25-54.

Kimball AB, Naegeli AN, Edson-Heredia E, Lin CY, Gaich C, Nikai E, Yosipovitch G. Psychometric properties of the Itch Numeric Rating Scale in patients with moderate to severe plaque psoriasis. *Br J Dermatol.* 2016;175(1):157-162.

Krakowski AC, Eichenfield LF, Dohil MA. Management of atopic dermatitis in the pediatric population. *Pediatrics*. 2008;122(4):812-824.

Kunz B, Oranje AP, Labrèze L, Stadler JF, Ring J, Taïeb A. Clinical validation and guidelines for the SCORAD index: consensus report of the European Task Force on Atopic Dermatitis. *Dermatology*. 1997;195(1):10-19.

Langenbruch A, Radtke M, Franzke N, Ring J, Foelster-Holst R, Augustin M. Quality of health care of atopic eczema in Germany: results of the national health care study AtopicHealth. *J Eur Acad Dermatol Venereol*. 2014;28(6):719-726.

Langley RG, Feldman SR, Nyirady J, van de Kerkhof P, Papavassilis C. The 5-point Investigator's Global Assessment (IGA) Scale: A modified tool for evaluating plaque psoriasis severity in clinical trials. *J Dermatolog Treat*. 2015;26(1):23-31.

Li M, Messaddeq N, Teletin M, Pasquali JL, Metzger D, Chambon P. Retinoid X receptor ablation in adult mouse keratinocytes generates an atopic dermatitis triggered by thymic stromal lymphopoietin. *Proc Natl Acad Sci U S A*. 2005;102(41):14795-14800.

Moniaga CS, Jeong SK, Egawa G, Nakajima S, Hara-Chikuma M, Jeon JE, Lee SH, Hibino T, Miyachi Y, Kabashima K. Protease activity enhances production of thymic stromal lymphopoietin and basophil accumulation in flaky tail mice. *Am J Pathol*. 2013;182(3):841-851.

Naegeli AN, Flood E, Tucker J, Devlen J, Edson-Heredia E. The Worst Itch Numeric Rating Scale for patients with moderate to severe plaque psoriasis or psoriatic arthritis. *Int J Dermatol.* 2015;54(6):715-722.

[NINDS] National Institute of Neurological Disorders and Stroke. User Manual for the Quality of Life in Neurological Disorders (Neuro-QoL) Measures, Version 2.0, March 2015.

Nomura T, Kabashima K. Advances in atopic dermatitis in 2015. *J Allergy Clin Immunol.* 2015;138(6):1548-1555.

Palmer CN, Irvine AD, Terron-Kwiatkowski A, Zhao Y, Liao H, Lee SP, Goudie DR, Sandilands A, Campbell LE, Smith FJ, O'Regan GM, Watson RM, Cecil JE, Bale SJ, Compton JG, DiGiovanna JJ, Fleckman P, Lewis-Jones S, Arseculeratne G, Sergeant A, Munro CS, El Houate B, McElreavey K, Halkjaer LB, Bisgaard H, Mukhopadhyay S, McLean WH. Common loss-of-function variants of the epidermal barrier protein filaggrin are a major predisposing factor for atopic dermatitis. *Nat Genet.* 2006;38:441-446.

Papp KA, Menter MA, Raman M, Disch D, Schlichting DE, Gaich C, Macias W, Zhang X, Janes JM. A randomized phase 2b trial of baricitinib, an oral Janus kinase (JAK) 1/JAK2 inhibitor, in patients with moderate to severe psoriasis. *Br J Dermatol*. 2016;174(6):1266-1276.

[PROMIS] Patient-Reported Outcomes Measurement Information System. Sleep-Related Impairment. 2015. Available at: http://www.healthmeasures.net/administrator/components/com_instruments/uploads/15-09-02_01-36-32_PROMISSleep-RelatedImpairmentScoringManual.pdf. Accessed May 31, 2018.

Reilly MC, Zbrozek AS, Dukes EM. The validity and reproducibility of a work productivity and activity impairment instrument. *Pharmacoeconomics*. 1993;4(5):353-365.

Schram ME, Spuls PI, Leeflang MM, Lindeboom R, Bos JD, Schmitt J. EASI, (objective) SCORAD and POEM for atopic eczema: responsiveness and minimal clinically important difference. *Allergy*. 2012;67(1):99-106.

Simpson EL. Comorbidity in atopic dermatitis. *Curr Dermatol Rep*. 2012;1(1):29-38.

Snaith RP. The Hospital Anxiety and Depression Scale. *Health and Quality of Life Outcomes*. 2003;1:29. Available at: http://www.hqlo.com/content/1/1/29. Accessed February 2017.

Stalder JF, Taïeb A, Atherton DJ, et al. for the European Task Force on Atopic Dermatitis. Severity scoring of atopic dermatitis: the SCORAD index. Consensus Report of the European Task Force on Atopic Dermatitis. *Dermatology*. 1993;186(1):23-31.

Taylor PC, Keystone EC, van der Heijde D, Weinblatt ME, del Carmen Morales L, Gonzaga JR, Yakushin S, Ishii T, Emoto K, Beattie S, Arora V, Gaich CL, Rooney T, Schlichting D, Macias WL, de Bono S, Tanaka. Baricitinib versus placebo or adalimumab in rheumatoid arthritis. *N Engl J Med.* 2017;376:652-662.

Tuttle KR, Brosius FC, Adler SG, Kretzler M, Mehta RL, Tumlin JA, Liu J, Silk ME, Cardillo TE, Duffin KL, Haas JV, Macias WL, Janes JM. Baricitinib in diabetic kidney disease: Results from a Phase 2, multicenter, randomized, double-blind, placebo-controlled study. Paper presented at: Annual Meeting of the American Diabetes Association; 5–9 June 2015; Boston, MA.

Vaddi K, Luchi M. JAK inhibition for the treatment of rheumatoid arthritis: a new era in oral DMARD therapy. *Expert Opin Investig Drugs*. 2012;21(7):961-973.

White D, Leach C, Sims R, Atkinson M, Cottrell D. Validation of the Hospital Anxiety and Depression Scale for use with adolescents. *Br J Psychiatry*. 1999;175(5):452-454.

Wilson SR, Thé L, Batia LM, Beattie K, Katibah GE, McClain SP, Pellegrino M, Estandian DM, Bautista DM. The epithelial cell-derived atopic dermatitis cytokine TSLP activates neurons to induce itch. *Cell*. 2013;155(2):285-295.

Yano C, Saeki H, Ishiji T, Ishiuji Y, Sato J, Tofuku Y, Nakagawa H. Impact of disease severity on sleep quality in Japanese patients with atopic dermatitis. *J Dermatol Sci*. 2013;72(2):195-197.

Ziegler SF, Roan F, Bell BD, Stoklasek TA, Kitajima M, Han H. The biology of thymic stromal lymphopoietin (TSLP). *Adv Pharmacol*. 2013;66:129-155.

Zigmond AS, Snaith RP. The Hospital Anxiety and Depression Scale. *Acta Psychiatr Scand*. 1983;67:361-370.

12. Appendices

Appendix 1. Abbreviations and Definitions

Term	Definition
AD	atopic dermatitis
ADSS	Atopic Dermatitis Sleep Scale
AE	adverse event: Any untoward medical occurrence in a patient or clinical investigation subject administered a pharmaceutical product that does not necessarily have a causal relationship with this treatment. An adverse event can therefore be any unfavorable and unintended sign (including an abnormal laboratory finding), symptom, or disease temporally associated with the use of a medicinal (investigational) product, whether or not related to the medicinal (investigational) product.
ALT	alanine aminotransferase
ALP	alkaline phosphatase
ANC	absolute neutrophil count
ANCOVA	analysis of covariance
AST	aspartate aminotransferase
BCG	Bacillus Calmette-Guérin
blinding/masking	A single-blind study is one in which the investigator and/or his staff are aware of the treatment but the patient is not, or vice versa, or when the sponsor is aware of the treatment but the investigator and/his staff and the patient are not. A double-blind study is one in which neither the patient nor any of the investigator or sponsor staff who are involved in the treatment or clinical evaluation of the subjects are aware of the treatment received.
BSA	body surface area
CI	confidence interval
C-SSRS	Columbia Suicide Severity Rating Scale
CSR	clinical study report
DLQI	Dermatology Life Quality Index
DMC	data monitoring committee
DNA	deoxyribonucleic acid
DVT	deep vein thrombosis
EASI	Eczema Area and Severity Index
ECG	electrocardiogram
eCOA	electronic clinical outcome assessment
eCRF	electronic case report form
eGFR	estimated glomerular filtration rate
enroll	The act of assigning a patient to a treatment. Patients who are enrolled in the trial are those who have been assigned to a treatment.
Enter	Patients entered into a trial are those who sign the informed consent form directly or through their legally acceptable representatives.
EQ-5D-5L	European Quality of Life–5 Dimensions–5 Levels
ERB	ethical review board
ETV	early termination visit
FDA	the Food and Drug Administration
FSH	follicle-stimulating hormone
GCP	good clinical practice

HADS	Hospital Anxiety Depression Scale
HBcAb	hepatitis B core antibody
HBsAg	hepatitis B surface antigen
HBV	hepatitis B virus
HCV	hepatitis C virus
HIV	human immunodeficiency virus
IB	Investigator's Brochure
ICF	informed consent form
ICH	International Council for Harmonisation
IGA	Investigator's Global Assessment
IL	interleukin
interim analysis	An interim analysis is an analysis of clinical study data, separated into treatment groups, that is conducted before the final reporting database is created/locked.
INR	international normalized ratio
Investigational product	A pharmaceutical form of an active ingredient or placebo being tested or used as a reference in a clinical trial, including products already on the market when used or assembled (formulated or packaged) in a way different from the authorized form, or marketed products used for an unauthorized indication, or marketed products used to gain further information about the authorized form.
IWRS	interactive web-response system
JAK	Janus kinase
LOCF	last observed carried forward
LSM	least squares mean
MACE	major adverse cardiovascular events
MI	myocardial infarction
MMRM	mixed-effects model of repeated measures
Neuro-QoL	Quality of Life in Neurological Disorders
NRI	nonresponder imputation
NRS	Numeric Rating Scale
PBI	Patient Benefit Index
PD	Pharmacodynamic(s)
PDE-4 inhibitor	phosphodiesterase type 4 inhibitor
PIQ	Pain Impact Questionnaire
PK	pharmacokinetic(s)
POEM	Patient-Oriented Eczema Measure
PPD	purified protein derivative
PRO/ePRO	patient-reported outcomes/electronic patient-reported outcomes
PROMIS	Patient-Reported Outcomes Measurement Information System
QD	once daily
QoL	quality of life
RA	rheumatoid arthritis
SAE	serious adverse event
SAP	statistical analysis plan
SCORAD	SCORing Atopic Dermatitis
STAT	signal transducer and activator of transcription
SUSAR	suspected unexpected serious adverse reaction
TB	tuberculosis
TBL	total bilirubin level
TCNI	topical calcineurin inhibitor

TCS	topical corticosteroids
TEAE	Treatment-emergent adverse event: An untoward medical occurrence that emerges during a defined treatment period, having been absent pretreatment, or worsens relative to the pretreatment state, which does not necessarily have to have a causal relationship with this treatment.
TSH	thyroid-stimulating hormone
TSLP	thymic stromal lymphopoietin
VAS	Visual Analog Scale
ULN	upper limit of normal
vIGA-AD	validated Investigator's Global Assessment for Atopic Dermatitis
VTE	venous thromboembolic event (deep vein thrombosis or pulmonary embolism)
WPAI-AD	The Work Productivity and Activity Impairment–Atopic Dermatitis

Appendix 2. Clinical Laboratory Tests

Hematology[a,b]	**Clinical Chemistry**[a,b]
Hemoglobin	**Serum Concentrations of:**
Hematocrit	Sodium
Erythrocyte count (RBC)	Potassium
Absolute Reticulocyte Count	Total bilirubin
Mean cell volume	Direct bilirubin
Mean cell hemoglobin	Alkaline phosphatase
Mean cell hemoglobin concentration	Alanine aminotransferase (ALT)
Leukocytes (WBC)	Aspartate aminotransferase (AST)
Platelets	Blood urea nitrogen (BUN)
Absolute counts of:	Creatinine
Neutrophils, segmented	Cystatin C
Neutrophils, juvenile (bands)	Uric acid
Lymphocytes	Calcium
Monocytes	Glucose
Eosinophils	Albumin
Basophils	Total protein
	Estimated glomerular filtration rate (eGFR)[e]
Urinalysis[a,b,c]	Creatine phosphokinase (CPK)
Color	
Specific gravity	**Other Tests**[a]
pH	Hepatitis B Surface antigen (HBsAg)[f]
Protein	Anti-Hepatitis B Core antibody (HBcAb)[f]
Glucose	HBV DNA[k]
Ketones	Anti-Hepatitis B Surface antibody (HBsAb)[f]
Bilirubin	Human immunodeficiency virus (HIV)[f]
Urobilinogen	Hepatitis C antibody[f,g]
Blood	Thyroid-stimulating hormone (TSH)
Leukocyte esterase	Exploratory storage samples (serum, plasma and mRNA)
Nitrite	Pregnancy Test[h]
	Follicle-stimulating hormone[f,i]
Lipids[a,d]	Serum immunoglobulin (IgA, IgG, IgM, and IgE)
Total cholesterol	QuantiFERON®-TB Gold or T-SPOT®.TB[j]
Low-density lipoprotein	PPD (local testing)
High-density lipoprotein	
Triglycerides	

Abbreviations: FSH = follicle-stimulating hormone; HBV = hepatitis B virus; Ig = immunoglobulin; mRNA = messenger ribonucleic acid; PPD = purified protein derivative; RBC = red blood cell; TB = tuberculosis; WBC = white blood cell.

a Assayed by sponsor-designated laboratory.

b Unscheduled or repeat blood chemistry, hematology, and urinalysis panels may be performed at the discretion of the investigator, as needed.

c Microscopic examination of sediment performed only if abnormalities are noted on the routine urinalysis.

d Fasting lipid profile. Patients should not eat or drink anything except water for 12 hours prior to test. If a patient attends these visits in a nonfasting state, this will not be considered a protocol violation.

[e] eGFR for serum creatinine calculated by the central laboratory using the Chronic Kidney Disease Epidemiology Collaboration (CKD-EPI) Creatinine 2009 equation.

[f] Test required at Visit 1 only to determine eligibility of patient for the study.

[g] A positive hepatitis C antibody (Hep C antibody) result will be confirmed with an alternate hepatitis C method.

[h] For all women of child-bearing potential, a serum pregnancy test will be performed at Visit 1 and a local urine pregnancy test will be performed at Visit 2 and at all subsequent study visits after Visit 3. If required per local regulations and/or institutional guidelines, pregnancy testing can occur at other times during the study treatment period.

[i] To confirm postmenopausal status for women ≥40 and <60 years of age who have had a cessation of menses, an FSH test will be performed. Non–child-bearing potential is defined as an FSH ≥40 mIU/mL and a cessation of menses for at least 12 months.

[j] The QuantiFERON®-TB Gold test is the preferred alternative to the PPD test for the evaluation of TB infection, and it may be used instead of the PPD test or T-SPOT®.TB test and may be read locally. If the QuantiFERON® TB Gold test is indeterminate, 1 retest is allowed. If the retest is indeterminate, then the patient is excluded from the study.

[k] HBV DNA testing will be done in those patients who are HBcAb+ at screening.

Appendix 3. Study Governance Considerations

Appendix 3.1. Regulatory and Ethical Considerations, Including the Informed Consent Process

Appendix 3.1.1. Informed Consent

The investigator is responsible for ensuring the following:

- that the patient understands the potential risks and benefits of participating in the study.
- that informed consent is given by each patient. This includes obtaining the appropriate signatures and dates on the informed consent form (ICF) prior to the performance of any protocol procedures and prior to the administration of investigational product.
- answering any questions the patient may have throughout the study and sharing in a timely manner any new information that may be relevant to the patient's willingness to continue his or her participation in the trial.

Appendix 3.1.2. Ethical Review

The investigator must give assurance that the ERB was properly constituted and convened as required by International Council for Harmonisation (ICH) guidelines and other applicable laws and regulations.

Documentation of ERB approval of the protocol and the ICF must be provided to Lilly before the study may begin at the investigative site(s). Lilly or its representatives must approve the ICF, including any changes made by the ERBs, before it is used at the investigative site(s). All ICFs must be compliant with the ICH guideline on Good Clinical Practice (GCP).

The study site's ERB(s) should be provided with the following:

- the current Investigator Brochure (IB) and updates during the course of the study
- informed consent form
- relevant curricula vitae

Appendix 3.1.3. Regulatory Considerations

This study will be conducted in accordance with:

- consensus ethics principles derived from international ethics guidelines, including the Declaration of Helsinki and Council for International Organizations of Medical Sciences (CIOMS) International Ethical Guidelines
- applicable ICH GCP Guidelines
- applicable laws and regulations

Some of the obligations of the sponsor will be assigned to a third party.

Appendix 3.1.4. Investigator Information

Physicians with a specialty in dermatology will participate as investigators in this clinical trial.

Appendix 3.1.5. Protocol Signatures

The sponsor's responsible medical officer will approve the protocol, confirming that, to the best of his or her knowledge, the protocol accurately describes the planned design and conduct of the study.

After reading the protocol, each principal investigator will sign the protocol signature page and send a copy of the signed page to a Lilly representative.

Appendix 3.1.6. Final Report Signature

Lilly will select a qualified investigator(s) from among investigators participating in the design, conduct, and/or analysis of the study to serve as the clinical study report (CSR) coordinating investigator. If this investigator is unable to fulfill this function, another investigator will be chosen by Lilly to serve as the CSR coordinating investigator.

The CSR coordinating investigator will sign the final CSR for this study, indicating agreement that, to the best of his or her knowledge, the report accurately describes the conduct and results of the study.

The sponsor's responsible medical officer and statistician will approve the final CSR for this study, confirming that, to the best of his or her knowledge, the report accurately describes the conduct and results of the study.

Appendix 3.2. Data Quality Assurance

To ensure accurate, complete, and reliable data, Lilly or its representatives will do the following:

- provide instructional material to the study sites, as appropriate
- provide sponsor start-up training to instruct the investigators and study coordinators. This training will give instruction on the protocol, the completion of the CRFs, and study procedures.
- make periodic visits to the study site
- be available for consultation and stay in contact with the study site personnel by mail, telephone, and/or fax
- review and evaluate CRF data and use standard computer edits to detect errors in data collection
- conduct a quality review of the database

In addition, Lilly or its representatives will periodically check a sample of the patient data recorded against source documents at the study site. The study may be audited by Lilly or its representatives and/or regulatory agencies at any time. Investigators will be given notice before an audit occurs.

The investigator will keep records of all original source data. This might include laboratory tests, medical records, and clinical notes. If requested, the investigator will provide the sponsor,

applicable regulatory agencies, and applicable ERBs with direct access to original source documents.

Appendix 3.2.1. Data Capture System

An electronic data capture system will be used in this study. The site maintains a separate source for the data entered by the site into the sponsor-provided electronic data capture system.

Electronic patient-reported outcome (ePRO) measures (e.g., a rating scale) and electronic clinical outcome assessments (eCOAs) are entered into an ePRO/eCOA instrument at the time that the information is obtained. In these instances where there is no prior written or electronic source data at the site, the ePRO/eCOA instrument record will serve as the source.

If ePRO/eCOA records are stored at a third-party site, investigator sites will have continuous access to the source documents during the study and will receive an archival copy at the end of the study for retention.

Any data for which the ePRO/eCOA instrument record will serve to collect source data will be identified and documented by each site in that site's study file.

Case report form data will be encoded and stored in InForm. Data managed by a central vendor, such as laboratory test data, will be stored electronically in the central vendor's database system. Data will subsequently be transferred from the central vendor to the Lilly data warehouse.

Data managed by a central vendor, such as laboratory test data, will be stored electronically in the central vendor's database system. Data will subsequently be transferred from the central vendor to the Lilly data warehouse.

Data from complaint forms submitted to Lilly will be encoded and stored in the global product complaint management system.

Appendix 3.3. Study and Site Closure

Appendix 3.3.1. Discontinuation of Study Sites

Study site participation may be discontinued if Lilly, the investigator, or the ERB of the study site judges it necessary for medical, safety, regulatory, or other reasons consistent with applicable laws, regulations, and GCP.

Appendix 3.3.2. Discontinuation of the Study

The study will be discontinued if Lilly judges it necessary for medical, safety, regulatory, or other reasons consistent with applicable laws, regulations, and GCP.

Appendix 4. Hepatic Monitoring Tests for Treatment-Emergent Abnormality

Selected tests may be obtained in the event of a treatment-emergent hepatic abnormality and may be required in follow-up with patients in consultation with the Lilly, its designee, or the clinical research physician.

Hepatic Monitoring Tests

Hepatic Hematology[a]	**Haptoglobin[a]**
Hemoglobin	
Hematocrit	**Hepatic Coagulation[a]**
RBC	Prothrombin Time
WBC	Prothrombin Time, INR
Neutrophils, segmented	
Lymphocytes	**Hepatic Serologies[a,b]**
Monocytes	Hepatitis A antibody, total
Eosinophils	Hepatitis A antibody, IgM
Basophils	Hepatitis B surface antigen
Platelets	Hepatitis B surface antibody
	Hepatitis B Core antibody
Hepatic Chemistry[a]	Hepatitis C antibody
Total bilirubin	Hepatitis E antibody, IgG
Direct bilirubin	Hepatitis E antibody, IgM
Alkaline phosphatase	
ALT	**Anti-nuclear antibody[a]**
AST	
GGT	**Alkaline Phosphatase Isoenzymes[a]**
CPK	
	Anti-smooth muscle antibody (or anti-actin antibody)[a]

Abbreviations: ALT = alanine aminotransferase; AST = aspirate aminotransferase; CPK = creatinine phosphokinase; GGT = gamma-glutamyl transferase; Ig = immunoglobulin; INR = international normalized ratio; RBC = red blood cell; WBC = white blood cell.

[a] Assayed by Lilly-designated or local laboratory.

[b] Reflex/confirmation dependent on regulatory requirements and/or testing availability.

Appendix 5. Liver Function Testing and Hepatic Safety Monitoring

Liver Function Testing and Hepatic Safety Monitoring

Analyte	Exclusion Criteria	Additional Hepatic Testing	Hepatic eCRF Reporting	Temporary Interruption of IP	Permanent Discontinuation of IP after Consultation with the Lilly-Designated Medical Monitor
Protocol Section	Section 6.2	Section 9.4.9	Section 9.4.9	Section 8.1.1	Section 8.1.2
ALT/AST	≥2x ULN	ALT ≥3x ULN	ALT ≥5x ULN on ≥2 consecutive tests	≥5x ULN	• >8x ULN • >5x ULN for >2 weeks • >3x ULN AND TBL >2x ULN or INR >1.5 • >3x ULN with symptoms[a]
ALP	≥2x ULN	≥2x ULN	≥2x ULN on ≥2 consecutive tests	No applicable criteria	• >3x ULN • >2.5x ULN AND TBL >2x ULN • >2.5x ULN with symptoms[a]
TBL	≥1.5x ULN	≥2x ULN	≥2x ULN (excluding Gilbert's syndrome)	No applicable criteria	• ALT or AST >3x ULN AND TBL >2x ULN • ALP >2.5x ULN AND TBL >2x ULN

Abbreviations: ALP = alkaline phosphatase; ALT = alanine aminotransferase; AST = aspartate aminotransferase; eCRF = electronic case report form; INR = international normalized ratio; IP = investigational product; TBL = total bilirubin level; ULN = upper level of normal.

[a] Fatigue, nausea, vomiting, right upper-quadrant pain or tenderness, fever, and/or rash

Appendix 6. Monitoring Tests for Confirmed VTE

Selected tests may be obtained in the event of a confirmed venous thromboembolic event (VTE) and may be required in follow-up with patients in consultation with Eli Lilly and Company, its designee, or the clinical research physician. The choice and optimal timing of these tests will be directed by the patient's management and may require ongoing follow-up after study discontinuation.

Protein C Functional
Protein S Clottable
Antithrombin III
APC Resistance
PT
APTT
Fibrinogen
Cardiolipin Antibodies
PT Gene
Factor VIII C Assay
Hexagonal Phase Phospholipid Neutralization
C-Reactive Protein
PTT Incubated Mixing
Dilute Russell Viper Venom
Platelet Neutralization
Factor V Leiden
MTHFR
Thrombin Time
Reptilase
Fibrinogen Antigen
Protein C Immunologic
Protein S Immunologic
Heparin fXa Inhibition

Abbreviations: APC = activated protein C; APTT = activated partial thromboplastin time; fXa = factor Xa; MTHFR = methylene tetrahydrofolate reductase; PT = prothrombin time; PTT = partial thromboplastin time.

Appendix 7. American Academy of Dermatology: Criteria for the Diagnosis and Assessment of Atopic Dermatitis

Features to be considered in diagnosis of patients with atopic dermatitis:

Essential Features—Must be present:

- pruritus
- eczema (acute, subacute, chronic)
    - typical morphology and age-specific patterns*
    - chronic or relapsing history

*Patterns include:

1)facial, neck, and extensor involvement in infants and children
2)current or previous flexural lesions in any age group
3)sparing of the groin and axillary regions

Important Features—Seen in most cases, adding support to the diagnosis:

- early age of onset
- atopy
    - personal and/or family history
    - Immunoglobulin E reactivity
- xerosis

Associated Features—These clinical associations help to suggest the diagnosis of atopic dermatitis but are too nonspecific to be used for defining or detecting atopic dermatitis for research and epidemiologic studies:

- atypical vascular responses (eg, facial pallor, white dermographism, delayed blanch response)
- keratosis pilaris/pityriasis alba/hyperlinear palms/ichthyosis
- ocular/periorbital changes
- other regional findings (e.g., perioral changes/periauricular lesions)
- perifollicular accentuation/lichenification/prurigo lesions

Exclusionary Features—It should be noted that a diagnosis of atopic dermatitis depends on excluding conditions, such as:

- scabies
- seborrheic dermatitis
- contact dermatitis (irritant or allergic)
- ichthyoses
- cutaneous T-cell lymphoma
- psoriasis
- photosensitivity dermatoses
- immune deficiency diseases
- erythroderma of other causes

Source: Eichenfield LF, Tom WL, Chamlin SL, Feldman SR, Hanifin JM, Simpson EL, Berger TG, Bergman JN, Cohen DE, Cooper KD, Cordoro KM, Davis DM, Krol A, Margolis DJ, Paller AS, Schwarzenberger K, Silverman RA, Williams HC, Elmets CA, Block J, Harrod CG, Smith Begolka W, Sidbury R. Guidelines of care for the management of atopic dermatitis: section 1. Diagnosis and assessment of atopic dermatitis. *J Am Acad Dermatol*. 2014;70(2):338-351.

Appendix 8. Classification of Potency for Topical Corticosteroids

Potency	Class	Topical Corticosteroid	Formulation
Ultra high	I	Clobetasol propionate	Cream 0.05%
		Diflorasone diacetate	Ointment 0.05%
High	II	Amcinonide	Ointment 0.1%
		Betamethasone dipropionate	Ointment 0.05%
		Desoximetasone	Cream or ointment 0.025%
		Fluocinonide	Cream, ointment or gel 0.05%
		Halcinonide	Cream 0.1%
	III	Betamethasone dipropionate	Cream 0.05%
		Betamethasone valerate	Ointment 0.1%
		Diflorasone diacetate	Cream 0.05%
		Triamcinolone acetonide	Ointment 0.1%
Moderate	IV	Desoximetasone	Cream 0.05%
		Fluocinolone acetonide	Ointment 0.025%
		Fludroxycortide	Ointment 0.05%
		Hydrocortisone valerate	Ointment 0.2%
		Triamcinolone acetonide	Cream 0.1%
	V	Betamethasone dipropionate	Lotion 0.02%
		Betamethasone valerate	Cream 0.1%
		Fluocinolone acetonide	Cream 0.025%
		Fludroxycortide	Cream 0.05%
		Hydrocortisone butyrate	Cream 0.1%
		Hydrocortisone valerate	Cream 0.2%
		Triamcinolone acetonide	Lotion 0.1%
Low	VI	Betamethasone valerate	Lotion 0.05%
		Desonide	Cream 0.05%
		Fluocinolone acetonide	Solution 0.01%
	VII	Dexamethasone sodium phosphate	Cream 0.1%
		Hydrocortisone	Lotion, cream, or ointment 2.5%
		Hydrocortisone acetate	Cream 1%
		Methylprednisolone acetate	Cream 0.25%

Source: [WHO] World Health Organization. Model Prescribing Information: Drugs used in skin diseases. 1997; Geneva.

Tadicherla S, Ross K, Shenefelt PD, Fenske NA. Topical corticosteroids in dermatology. *J Drugs Dermatol.* 2009;8(12):1093-1105.

Appendix 9. Protocol Amendment I4V-MC-JAIY(a) Summary A Multicenter, Randomized, Double-Blind, Placebo-Controlled, Phase 3 Study to Evaluate the Efficacy and Safety of Baricitinib in Combination with Topical Corticosteroids in Adult Patients with Moderate to Severe Atopic Dermatitis

Overview

Protocol I4V-MC-JAIY A Multicenter, Randomized, Double-Blind, Placebo-Controlled, Phase 3 Study to Evaluate the Efficacy and Safety of Baricitinib in Combination with Topical Corticosteroids in Adult Patients with Moderate to Severe Atopic Dermatitis has been amended. The new protocol is indicated by amendment (a) and will be used to conduct the study in place of any preceding version of the protocol.

Amendment Summary for Protocol I4V-MC-JAIY Amendment (a)

Section #	Description of Change	Brief Rationale
Section 5.1.2	crisaborole was listed in the description of rescue therapies in Period 2. This was an error, crisaborole is not considered a rescue therapy and is allowed during the trial (Permitted Medications Section 7.7.1).	Minor error was corrected
Section 5.4	in the scientific rationale for study design section, intolerance to existing topical therapies was listed. This was an error and was corrected to only include inadequate response to topic therapies	Minor error was corrected
Section 6.1	leukotriene inhibitors were removed from washout	Leukotriene inhibitors are not a prohibited medication because evidence suggests little impact to AD, so no need for washout
Section 7.7.4	wording on rescue with high and ultra-high potency TCS use was updated	Modified for clarity
Section 8.1.2 and Appendix 5	Eosinophilia (>5%) was removed from permanent discontinuation of IP criteria	Elevated eosinophils are very common in patients with moderate to severe AD and do not reflect an increased risk for liver events.

Revised Protocol Sections

Note:	Deletions have been identified by ~~strikethroughs~~. Additions have been identified by the use of underscore.

5.1.2. Period 2: Double–Blind, Placebo-Controlled Treatment

Patient will be randomized at a 1:1:1 ratio into 1 of the 3 treatment groups (placebo QD, baricitinib 2-mg QD, or baricitinib 4-mg QD). Investigational product will be administered daily for 16 weeks (treatment period Visits 2 through 8; Section 7). All patients will be required to use emollients daily. Daily diaries will continue to be utilized throughout the treatment period. Download of this data will be required at study visits. TCS will be dispensed at V2 and used on affected areas as described in section 7.7.2. Topical calcineurin inhibitors (TCNIs) is also allowed, but TCNI use should be limited to problem areas (e.g. face and skin folds). The use of higher potency TCS~~, crisaborole,~~ and systemic therapies for the treatment of AD are not allowed, except as part of rescue therapy for patients not responding to treatment. Details of background topical therapy, as well as rescue therapy and rescue criteria are included in Section 7.7. Assessments of disease severity will be performed by the investigator at all study visits including unscheduled and early termination visits (ETVs).

5.4. Scientific Rationale for Study Design

This study will enroll moderate to severe AD patients with a history of inadequate response ~~or intolerance~~ to existing topical therapies for whom a systemic treatment such as baricitinib may therefore be appropriate.

6.1. Inclusion Criteria

Type of Patient and Disease Characteristics

[6] agree to discontinue use of the following excluded medications/treatments for at least 4 weeks prior to randomization (Visit 2) and throughout the study:

a. oral systemic corticosteroids ~~and leukotriene inhibitors~~

b. systemic immunomodulators, including, but not limited to, cyclosporine, methotrexate, mycophenolate mofetil, and azathioprine

c. any other systemic therapy used to treat AD or symptoms of AD (approved or off-label use)

7.7.4. Rescue Therapy

Rescue with High- and Ultra-High-Potency TCS

High- or ultra-high-potency TCS may be used once daily for up to 14 consecutive days or less, or based on the maximum duration recommended in the prescribing information.

8.1.2. Permanent Discontinuation from Investigational Product

Investigational product should be permanently discontinued if the patient requests to discontinue investigational product.

Discontinuation of the investigational product for abnormal liver tests should be considered by the investigator when a patient meets 1 of the following conditions after consultation with the Lilly-designated medical monitor:

- ALT or AST >8x ULN
- ALT or AST >5x ULN for more than 2 weeks
- ALT or AST >3x ULN and total bilirubin level (TBL) >2x ULN or international normalized ratio (INR) >1.5
- ALT or AST >3x ULN with the appearance of fatigue, nausea, vomiting, right upper-quadrant pain or tenderness, fever, and/or rash~~, and/or eosinophilia (>5%)~~
- ALP >3x ULN
- ALP >2.5x ULN and TBL >2x ULN
- ALP >2.5x ULN with the appearance of fatigue, nausea, vomiting, right quadrant pain or tenderness, fever, and/or rash~~, and/or eosinophilia (>5%)~~

NOTE: Patients who are discontinued from investigational product due to a hepatic event or liver test abnormality should have additional hepatic safety data collected via the hepatic safety eCRF.

Appendix 5. Liver Function Testing and Hepatic Safety Monitoring

Table footnotes:

[a] Fatigue, nausea, vomiting, right upper-quadrant pain or tenderness, fever, and/or rash~~, and/or eosinophilia (>5%).~~

Leo Document ID = 0f04528a-106c-4e12-ae0b-97e05bca33ac

Approver:PPD
Approval Date & Time: 18-Dec-2018 16:23:01 GMT
Signature meaning: Approved

Approver: PPD
Approval Date & Time: 18-Dec-2018 16:51:32 GMT
Signature meaning: Approved

1. Statistical Analysis Plan for US and Japan: I4V-MC-JAIY (a): A Multicenter, Randomized, Double-Blind, Placebo-Controlled, Phase 3 Study to Evaluate the Efficacy and Safety of Baricitinib in Combination with Topical Corticosteroids in Adult Patients with Moderate to Severe Atopic Dermatitis

BREEZE-AD7

Baricitinib (LY3009104) Atopic Dermatitis

Study I4V-MC-JAIY is a Phase 3, multicenter, randomized, double-blind, placebo-controlled, parallel-group, outpatient study evaluating the efficacy and safety of baricitinib 4 mg once daily (QD) plus topical corticosteroids (TCS) and 2 mg QD plus TCS, as compared to placebo plus TCS in adult patients with moderate to severe atopic dermatitis.

Eli Lilly and Company
Indianapolis, Indiana USA 46285
Protocol I4V-MC-JAIY
Phase 3
Statistical Analysis Plan Version 1 electronically signed and approved by Lilly:
01 April 01, 2019
Statistical Analysis Plan Version 2 electronically signed and approved by Lilly on date provided below.

Approval Date: 12-Aug-2019 GMT

2. Table of Contents

Section **Page**

Table of Contents

Table **Page**

Table of Contents

Figure **Page**

3. Revision History

Statistical Analysis Plan (SAP) Version 1 is based on Protocol I4V-MC-JAIY(a). Statistical Analysis Plan Version 2 was approved prior to unblinding. A summary of changes between Version 1 and Version 2 are as follows:

Section	Summary of Changes
4.3. Exploratory objectives	• Updated the time frame for the objective of frequency of patient-reported "no itch" and "no pain" by starting from Week 0. • Added new exploratory analyses on Hospital Anxiety Depression (HADS), Dermatology Life Quality Index (DLQI), Atopic Dermatitis Sleep Scale (ADSS), and Patient-Oriented Eczema Measure (POEM). • Added new exploratory analyses on early responders.
6.11.1. Background TCS	• Updated the weight of dispensed topical corticosteroids (TCS) tube to align with JAHL SAP.
6.2.2. Definition on Baseline and Postbaseline Measures	• Updated the derivations of itch weekly scores.
6.2.3. Analysis Methods	• Updated the structures of covariance matrix in mixed model repeated measures (MMRM) model.
6.6. Multiple Comparisons	• Updated the graphical testing procedure.
6.16 Subgroup Analysis	• Added subgroup analyses for East Asia vs. other.
Table 6.6	• Updated the endpoints on the number days of itch-free and pain-free days.
Table 6.7	• Updated analyses on the number of itch-free and pain-free days. • Change the type of analysis of covariance (ANCOVA) (modified last observation carried forward [mLOCF]) analyses on PROMIS from "exploratory analyses" to "sensitivity analyses."

4. Study Objectives

4.1. Primary Objective

The primary objective of this study is to test the hypothesis that baricitinib 4 mg once daily (QD) plus topical corticosteroids (TCS) or baricitinib 2 mg QD plus TCS is superior to placebo plus TCS in the treatment of patients with moderate to severe atopic dermatitis (AD), as assessed by the proportion of patients achieving the validated Investigator's Global Assessment for AD (vIGA-AD, referred to throughout the SAP as IGA) of 0 or 1 with a ≥2-point improvement at Week 16.

In particular, the associated estimand for this objective is to measure the effect of therapy with baricitinib as assessed by the proportion of patients with a response of IGA 0 or 1 at Week 16 assuming treatment response disappears after patients are rescued or discontinue from the study or treatment. See Sections 6.4.1 and 6.12.1 on how this estimand handles outcomes after the occurrence of any intercurrent event through nonresponder imputation (NRI).

4.2. Secondary Objectives

4.2.1. Key Secondary Objectives

These are prespecified objectives that will be adjusted for multiplicity.

Objectives	Endpoints
To compare the efficacy of baricitinib 2 mg QD + TCS or baricitinib 4 mg QD + TCS to placebo + TCS in AD during the 16-week double-blind placebo-controlled treatment period as measured by improvement in signs and symptoms of AD.	• Proportion of patients achieving EASI75 at 16 weeks • Proportion of patients achieving EASI90 at 16 weeks • Percent change from baseline in EASI score at 16 weeks • Proportion of patients achieving SCORAD75 at 16 weeks
To compare the efficacy of baricitinib 2 mg QD + TCS or baricitinib 4 mg QD + TCS to placebo + TCS in AD during the 16-week double-blind placebo-controlled treatment period as assessed by patient-reported outcome measures.	• Proportion of patients achieving a 4-point improvement from baseline in Itch NRS at 2 days, 1 week, 2 weeks, 4 weeks, and 16 weeks • Mean change from baseline in the score of Item 2 of the ADSS at 1 week and 16 weeks • Mean change from baseline in Skin Pain NRS at 16 weeks

4.2.2. Other Secondary Objectives

These are prespecified objectives that will not be adjusted for multiplicity.

Objectives	Endpoints
To test the hypothesis that baricitinib 2 mg QD + TCS or baricitinib 4 mg QD + TCS is superior to placebo + TCS in the treatment of patients with moderate to severe AD.	• Proportion of patients achieving IGA of 0 or 1 with a ≥2-point improvement at Week 4
To compare the efficacy of baricitinib 2 mg QD + TCS or baricitinib 4 mg QD + TCS to placebo + TCS in AD during the 16-week double-blind placebo-controlled period as measured by signs and symptoms of AD.	• Proportion of patients achieving EASI50 at 16 weeks • Proportion of patients achieving IGA of 0 at 16 weeks • Mean change from baseline in SCORAD at 16 weeks • Proportion of patients achieving SCORAD90 at 16 weeks • Mean change from baseline in BSA affected at 16 weeks • Proportion of patients developing skin infections requiring antibiotic treatment by Week 16 • Mean gram quantity of background TCS used over 16 weeks (tube weights)
To compare the efficacy of baricitinib 2 mg QD + TCS or baricitinib 4 mg QD + TCS to placebo + TCS in AD during the 16-week, double-blind, placebo-controlled treatment period as assessed by patient-reported outcome/QoL measures.	• Percent change from baseline in Itch NRS at 2 days, 1 week, 4 weeks, and 16 weeks • Mean change from baseline in Itch NRS at 2 days, 1 week, 4 weeks and 16 weeks • Proportion of patients achieving a 4-point improvement from baseline in Skin Pain NRS at 16 weeks • Mean change from baseline in the total score of the POEM at 16 weeks • Mean change in the PGI-S-AD scores at 16 weeks • Mean change from baseline in HADS at 16 weeks • Mean change in the DLQI scores at 16 weeks • Mean change in the WPAI scores at 16 weeks • Mean change in the EQ-5D-5L scores at 16 weeks • Mean number of days without use of background TCS over 16 weeks

4.3. Exploratory Objectives

The exploratory objectives of this study are the following:

Objectives/Endpoints
• Frequency of patient-reported "no itch" (Itch NRS score = 0) days from daily diaries from Week 0 to Week 16 • Frequency of patient-reported "no pain" (Skin Pain NRS score = 0) days from daily diaries from Week 0 to Week 16 • Mean change from baseline in PIQ – Itch Interference score • Mean change from baseline in PIQ – Activity and Clothing score • Mean change from baseline in PIQ – Mood and Sleep score • Mean change from baseline in PIQ – Scratching Behavior score • Mean change from baseline in PROMIS – Sleep-Related Impairment score • Mean change from baseline in Neuro-QoL – Cognitive Function score • Patient Benefit Index score at 16 weeks – global score plus the following subscales: o Reducing social impairments o Reducing psychological impairments o Reducing impairments due to therapy o Reducing physical impairments o Having confidence in healing • Proportion of patients achieving PBI global score ≥1 at 16 weeks • Mean change from baseline in the score of Item 1 of the ADSS at 1 week and 16 weeks • Proportion of patients achieving a ≥1-point improvement in the score of Item 1 of the ADSS for those with baseline Item 1 score ≥1 • Proportion of patients achieving a ≥1-point improvement in the score of Item 2 of the ADSS for those with baseline Item 2 score ≥ 1 • Proportion of patients achieving a ≥2-point improvement in the score of Item 2 of the ADSS for those with baseline Item 2 score ≥2 • Mean change from baseline in the score of Item 3 of the ADSS at 1 week and 16 weeks • Proportion of patients achieving a ≥1-point improvement in the score of Item 3 of the ADSS for those with baseline Item 3 score ≥1 • To evaluate changes from baseline in immunoglobulin E levels during the study • To evaluate changes from baseline in eosinophil levels during the study • To assess time to 4-point Itch NRS improvement during the first 14 days after initiation of treatment • To assess time to 4-point improvement in Skin Pain during the first 14 days after the initiation of treatment • Proportion of patients achieving a ≥4-point improvement in DLQI total score for those with baseline DLQI total score ≥4 • Proportion of patients achieving DLQI total score 0 or 1 • Proportion of patients achieving a ≥ 4-point improvement in POEM total score for those with baseline total score ≥4 • Proportion of patients achieving HADS Anxiety Score <8 for those with baseline HADS Anxiety Score ≥8 • Proportion of patients achieving HADS Depression Score <8 for those with baseline HADS Depression Score ≥8 • Proportion of patients achieving improvement with HADS Anxiety Score or HADS Depression Score <8 for those with baseline HADS Anxiety Score ≥8 or HADS Depression Score ≥8 • Mean change from baseline in HADS subscale scores

- Proportion of patients achieving ≥4-point improvement in Itch NRS or IGA≤1 at Week 16 for those with ≥3-point improvement in Itch NRS or IGA≤2 at Week 4
- Proportion of patients achieving ≥ 4-point improvement in Itch NRS or IGA≤1 at any time between Week 8 and Week 16 for those with ≥3-point improvement in Itch NRS or IGA≤2 at Week 4
- Proportion of patients achieving [≥4-point improvement in Itch NRS and IGA≤2] or [IGA≤1] at Week 16 for those with ≥3-point improvement in Itch NRS or IGA≤2 at Week 4
- Proportion of patients achieving [≥4-point improvement in Itch NRS and IGA≤2] or [IGA≤1] at any time between Week 8 and Week 16 for those with ≥3-point improvement in Itch NRS or IGA≤2 at Week 4

5. Study Design

5.1. Summary of Study Design

Study I4V-MC-JAIY (JAIY) is a Phase 3, multicenter, randomized, double-blind, placebo-controlled, parallel-group, outpatient study evaluating the efficacy and safety of baricitinib 2 mg QD and 4 mg QD, in combination with TCS, as compared to placebo in combination with TCS, in adult patients with moderate to severe AD. The study is divided into 3 periods, a 5-week Screening period, a 16-week Double-Blinded Treatment period, and a 4-week Post-Treatment Follow-Up period. For those patients who complete the 16-week treatment period, there is an option to participate in the long-term extension study I4V-MC-JAHN (JAHN).

Approximately 300 patients ≥18 years of age who have responded inadequately to topical therapy will be randomized at a 1:1:1 ratio to receive placebo QD, baricitinib 2 mg QD, or baricitinib 4 mg QD in combination with TCS (100 patients in each treatment group). Patients will be stratified at randomization according to disease severity (Investigator's Global Assessment [IGA] 3 vs. 4) and geographic region.

Study JAIY will consist of 3 periods:

- Period 1: Screening period is between 8 and 35 days prior to Week 0 (Visit 2)
- Period 2: Double-Blind, Placebo-Controlled Treatment period from Week 0 (Visit 2) through Week 16 (Visit 8)
- Period 3: Post-Treatment Follow-Up period from last treatment visit at Week 16 (Visit 8) or Early Termination Visit (ETV) to approximately 28 days after the last dose of investigational product

Figure JAIY.5.1 illustrates the study design. The blinding procedure is described in the Protocol.

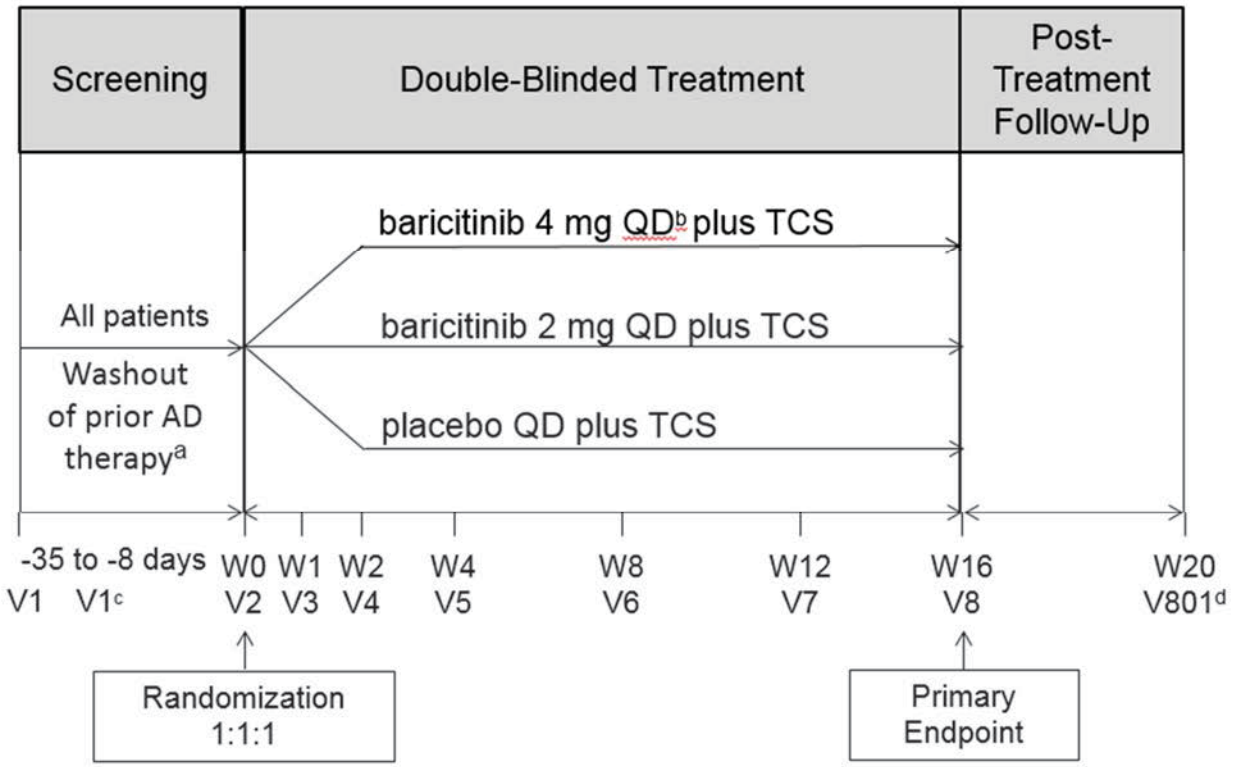


Figure JAIY.5.1. Illustration of study design for Clinical Protocol I4V-MC-JAIY.

Abbreviations: AD = atopic dermatitis; eGFR = estimated glomerular filtration rate; PPD = purified protein derivative; QD = once daily; TCS = topical corticosteroids; V = visit; W = week.

[a] Applicable to patients taking topical treatments (excluding emollients) or systemic treatments for AD at the time of screening.

[b] For patients randomized to the 4-mg once daily dose who have renal impairment (defined as eGFR <60 mL/min/1.73 m^2), the baricitinib dose will be 2 mg once daily.

[c] Patients for whom PPD skin test for the evaluation of tuberculosis infection was performed at V1 must return and PPD test must be read 48 to 72 hours after Visit 1 (post-PPD).

[d] Occurs approximately 28 days after the last dose of investigational product. Not required for those patients entering the long-term extension study JAHN.

5.2. Method of Assignment to Treatment

Patients who meet all criteria for enrollment will be randomized in a 1:1:1 ratio (placebo; baricitinib 2 mg; baricitinib 4 mg) to double-blind treatment at Visit 2 (Week 0). Randomization will be stratified by geographic region (Europe [EU], Japan [JPN], rest-of-world [ROW]) and disease severity at baseline (IGA 3 vs. 4). Assignment to treatment groups will be determined by a computer-generated random sequence using an IWRS. The IWRS will be used to assign packages containing double-blind investigational product tablets to each patient according to the study schedule of activities. Site personnel will confirm that they have located the correct packages by entering a confirmation number found on the packages into the IWRS.

This study will be conducted internationally in multiple sites. Table JAIY.5.1 describes how regions will be defined for stratification. Regions may be combined for statistical analyses in the

case when one of the region strata fails to meet the required minimum number of 30 patients. The 2 region strata with the least number of patients will then be pooled.

Table JAIY.5.1. Geographic Regions for Stratification

Region	Countries
Europe	Germany, Italy, Poland, Spain, Austria
Japan	Japan
Rest of World	Taiwan, Australia, Korea, Argentina

6. A Priori Statistical Methods

6.1. Determination of Sample Size

Study JAIY will aim to enroll approximately 300 patients ≥18 years of age. The proposed sample size will ensure a 89% power to detect an absolute difference of 20% between the baricitinib 4 mg and placebo treatment groups and the baricitinib 2 mg and placebo treatment groups, each using a 2-sided alpha of 0.025 and a Fisher´s exact test, assuming a 10% placebo response rate for the primary endpoint. The assumptions are based on what was observed in the Phase 2 study (JAHG). The proposed end point of IGA 0 or 1 represents patients whose AD is clear or almost clear from a baseline of moderate or severe disease. The anticipated effect size represents 2 times more patients achieving this benefit compared to placebo, which, in discussion with therapeutic experts, is of a magnitude that is considered clinically relevant.

Sample size and power estimates were obtained from nQuery® Advisor 7.0.

6.2. General Considerations

This plan describes *a priori* statistical analyses for efficacy, health outcomes, and safety that will be performed.

Statistical analysis of this study will be the responsibility of Eli Lilly and Company (Lilly). The statistical analyses will be performed using SAS® Version 9.4 or higher.

Not all displays described in this SAP will necessarily be included in the clinical study report (CSR). Not all displays will necessarily be created as a "static" display. Some may be incorporated into interactive display tools instead of or in addition to a static display. Any display described in this SAP and not included in the CSR will be available upon request.

Statistical tests of treatment effects and confidence intervals (CIs) will be performed at a 2-sided significance level of 0.05, unless otherwise stated (e.g., graphical multiple testing strategy in Section 6.6).

Data collected at ETVs will be mapped to the closest scheduled visit number for that patient if it falls within the visit window as discussed in Section 6.2.2. For by-visit summaries, only visits in which a measure was scheduled to be collected will be summarized. Any unscheduled visit data will be included at the patient-level listings. However, the data will still be used in other analyses, including shift analyses for safety analytes, change from baseline using modified last observation carried forward (mLOCF) for efficacy analyses, and other categorical analyses including safety.

6.2.1. Analysis Populations

Intent-to-treat (ITT) population: The ITT population analysis set is defined as all randomized patients.

Per-protocol Set (PPS): PP subset of the ITT analysis set will include those patients who do not have any significant or important protocol violations. Qualifications for and identification of

significant or important protocol violations will be determined while the study remains blinded, prior to database lock.

Follow-up population: The follow-up population is defined as patients who entered the follow-up period.

Unless otherwise specified, the efficacy and health outcome analyses will be conducted on the ITT population (Gillings and Koch 1991), which seeks to preserve the benefits of randomization and avoid selection bias. Patients will be analyzed according to the treatment to which they were randomized. In addition, the analyses of primary and key secondary endpoints will be repeated using the PPS population.

Safety population: The safety population is defined as all randomized patients who receive at least 1 dose of investigational product and who did not discontinue from the study for the reason 'Lost to Follow-up' at the first postbaseline visit.

Safety analyses will be performed using the safety population. Patients will be analyzed according to the treatment regimen to which they were assigned. Analyses of the safety endpoints, many of which are incidence based, will include all patients in the safety population, unless specifically stated otherwise.

In the rare situation where a patient is Lost to Follow-up at the first postbaseline visit, but some safety data exists (e.g., unscheduled laboratory assessments) after first dose of study drug, a listing of the data or a patient profile may be provided, when requested.

6.2.2. Definition of Baseline and Postbaseline Measures

The baseline value for efficacy and health outcomes variables measured at scheduled visits is defined as the last non-missing measurement on or prior to the date of first study drug administration (expected at Week 0, Visit 2).

The baseline value for the daily diary assessments (Itch NRS, ADSS, Skin Pain NRS, PGI-S-AD) is the mean of the non-missing assessments in the 7 days prior to the date of first study drug administration (expected at Week 0, Visit 2).

If there are less than 4 non-missing assessments in the baseline diary window, the interval lower bound can be extended up to 7 additional days, one day at a time, to obtain the most recent 4 non-missing values. If there are not at least 4 non-missing assessments in the baseline period, the baseline mean is missing.

Baseline for the safety analyses is defined as the last non-missing scheduled (planned) measurement on or prior to the date of first study drug administration for continuous measures by-visit analyses and all non-missing measurements on or prior to the date of first study drug administration for all other analyses.

Postbaseline measurements are collected after study drug administration through Week 16 (Visit 8) or early discontinuation visit. Efficacy data collected at scheduled visits (e.g., eCOA, ClinRO) will be used in all analyses unless it is missing. If an assessment is missing at a scheduled visit, an unscheduled post-baseline assessment can be used provided it falls within the

window interval as follows: a ±2 day window is used for Visit 3 (Week 1), Visit 4 (Week 2), and Visit 5 (Week 4); and a ±4 day window is used to Visit 6 (Week 8), Visit 7 (Week 12), and Visit 8 (Week 16). If there is more than 1 unscheduled visit within the defined visit window and no scheduled visit assessment is available, the unscheduled visit closest to the scheduled visit date will be used. If two unscheduled visits of equal distance are available, then the latter of the two will be used.

Postbaseline daily diary endpoints will be the mean of weekly visit windows (diary windows) anchored on day of first dose (Day 1) for Weeks 1 through 14 as follows: Week 1 (Days 1 through7), Week 2 (Days 8 through 14), Week 3 (Days 15 through 21), …, Week 14 (Days 92 through 98).

Week 16 Daily Diary Window Construction

The following sequential steps will be used to determine the Week 16 diary window. The general goal is to anchor on the scheduled Week 16 visit (or a proximal unscheduled visit) if such a visit exists or to use an interval based on days in study for cases where a scheduled Week 16 or a proximal surrogate does not exist.

Step 1: If the Week 16 scheduled visit exists, the Week 16 diary interval is the 7 days prior to the Week 16 date provided that window has at least 4 non-missing observations. If there are less than 4 non-missing observations, the diary window's lower bound will be extended 1 day at a time (up to day 99) to a maximum of 14 days prior to the Week 16 date until 4 non-missing observations are obtained. If, after extending this diary window's lower bound to 14 days, there are less than 4 non-missing observations then go to Step 2.

Step 2: If the Week 16 scheduled visit does not exist, the 7 days prior to the last visit (scheduled or unscheduled) occurring after Day 105, will constitute the Week 16 diary window provided that window contains at least 4 non-missing observations. If there are less than 4 non-missing observations, the diary window's lower bound will be extended 1 day at a time (up to Day 99) to a maximum of 14 days prior to the unscheduled visit date until 4 non-missing observations are obtained. If, after extending this diary window's lower bound to 14 days, there are less than 4 non-missing observations then go to Step 3.

Step 3: If neither a Week 16 scheduled visit is available nor an unscheduled visit to act as a surrogate for the Week 16 diary window, then the Week 16 window will be Day 106 to Day 112. If there are less than 4 non-missing observations, the dairy window's lower bound will be extended 1 day at a time to Day 99 until 4 non-missing observations are obtained.

If the steps above do not detect a window with at least 4 non-missing observations, then the Week 16 window is 7 days from either the Week 16 visit, the surrogate visit or Days 106 through 112 and the mean is missing and subject to imputation rules.

Week 15 Daily Diary Window Construction

The lower boundary of the Week 15 diary window is defined as Day 99. The upper bound of the Week 15 diary window is the minimum of either Day 105 or the lower bound of the Week 16

diary window -1. Consequently, Week 15 may be less than 4 days if the Week 16 scheduled visit is before Day 112. Moreover, as Week 15 diary window cannot exceed 7 days, there could be daily assessments between Weeks 15 and 16 diary windows that do not fall into a diary window. If after constructing the diary windows, there are fewer than 4 non-missing values the mean for Week 15 is missing and subject to imputation rules.

Handling of Duplicate Diary Records

If there is more than one diary record on a particular date, the first record on that particular date will be used in the analysis.

Note, as some analyses require use of the primary censoring rule, assessments collected on the day of rescue or afterwards will be excluded from the weekly visit interval calculation when implementing the rule for daily diary. If, after exclusion of these records, there are less than 4 non-missing assessments, the weekly interval which implements the primary censoring rule will be missing. The post-study follow-up weekly score for daily diaries will be calculated as the mean of the 7 days prior to the follow-up visit which occur after last dose of study treatment.

Postbaseline measures for the safety analyses are defined as the non-missing scheduled (planned) measurements after the date of first study drug administration for continuous measures by-visit analyses and all non-missing measurements after the date of first study drug administration for all other analyses.

6.2.3. Analysis Methods

The main analysis method of categorical efficacy variables and health outcomes variables will use a logistic regression analysis with region, baseline disease severity (IGA), baseline value and treatment group in the model, except for the analysis on IGA. For IGA, the logistic regression model will include region, baseline disease severity (IGA), and treatment group. Firth's correction will be used in order to accommodate (potential) sparse response rates. The p-value for the odds ratio from the logistic regression model will be used for statistical inference, unless Firth's correction still results in quasi-separation. In that case, Fisher's exact test will be used for statistical inference. The difference in percentages and 100(1-alpha)% CI of the difference in percentages using the Newcombe-Wilson method without continuity correction will be reported. The p-value from the Fisher's exact test will also be produced as a secondary analysis.

The main analysis method for all continuous efficacy and health outcomes variables will use mixed model repeated measures (MMRM) analysis. The MMRM model will use a restricted maximum likelihood (REML) estimation. The model will include treatment, region, baseline disease severity (IGA), visit, and treatment-by-visit-interaction as fixed categorical effects and baseline and baseline-by-visit-interaction as fixed continuous effects. For daily diary assessments, the model for analyses up to Week 16 will include all weekly assessments. An unstructured (co)variance structure will be used to model the between- and within-patient errors. If this analysis fails to converge, the heterogeneous autoregressive [ARH(1)], followed by the heterogeneous compound symmetry (CSH), followed by the heterogeneous Toeplitz (TOEPH), followed by autoregressive [AR(1)], followed by compound symmetry (CS) will be used. The

Kenward-Roger method will be used to estimate the degrees of freedom. Treatment least squares means (LSM) will be estimated within the framework of the MMRM using type 3 sums of squares. Differences in LSM between each dose of baricitinib and placebo (and associated p-values, standard errors and 95% CI) will be used for statistical inference. The LSM difference, standard error, p-value and 95% CI will be reported.

Treatment comparisons for continuous efficacy and health outcomes variables may also be made using analysis of covariance (ANCOVA) for primary and key secondary objectives. When an ANCOVA model is used, the model includes region, baseline disease severity, treatment group, and baseline value. Treatment LSM will be estimated within the framework of the ANCOVA using type 3 sums of squares. Reported differences in LSM and associated p-values, standard errors and 95% CI will be used for statistical inference. Treatment-by-region interaction will also be added to the model for sensitivity purposes and is discussed in Section 6.5.

Beginning on Day 14 a Cox proportional hazard (CPH) model of time (in days) to first observance of a 4-point itch reduction with effects for treatment, region, baseline mean itch, and disease severity will be used to test for treatment differences from placebo. For this analysis, daily itch scores will be compared to the baseline to determine if a 4-point itch reduction has been achieved in patients with a baseline itch of at least 4. The baseline for Itch NRS is defined in Section 6.2.2. Beginning on Day 14, the day on which the first time itch NRS is reduced by at least 4 will be modeled. If any significant difference between any baricitinib dose and placebo is observed then the same analyses will be run on Day 13. This process of evaluating at the next lowest day will proceed until no significant differences are observed. No adjustments for multiple tests and multiple comparisons will be used. This analysis uses the Primary Censoring Rule for patients who are rescued or permanently discontinue study drug (see Section 6.4). Missing daily itch data will be replaced using NRI rule which means missing data is replaced with a non-response which would entail replacing missing values with a time to event of >14 days, censored; >13 days, censored, etc., depending on the window being used. If the model assumptions for the CPH model do not hold, a log-rank test will be used.

Restricted mean survival time will be evaluated at Days 2, 3, 4 and 5 as an exploratory analysis. The model will include terms for baseline disease severity (IGA), baseline itch, region, and treatment group.

Fisher's exact test will be used to test for differences between each baricitinib dose and placebo in proportions of patients experiencing adverse events (AEs), discontinuation from study drug, and for other categorical safety data. Continuous vital signs, body weight, and other continuous safety variables, including laboratory variables will be analyzed by an ANCOVA with treatment group and baseline value in the model. The significance of within-treatment group changes from baseline will be evaluated by testing whether or not the treatment group LSM changes from baseline are different from zero; the standard error for the LSM change will also be displayed. Differences in LSM will be displayed, with the p-value associated with the LSM comparison to placebo and a 95% CI on the LSM difference also provided. In addition to the LSMs for each group, the within-group p-value for the change from baseline will be displayed.

Confusion table will be used to calculate the proportions of responder patients defined at week 16, and at any time between Weeks 8 and 16, for those with improvement on itch NRS or IGA at early week. The confusion table will generate the Positive Prediction Value (PPV) and Negative Prediction Value (NPV), which can access the performance of predicting late week responders using early week data. All missing values will be treated by NRI method.

6.2.4. Derived Data

- Age (year), derived using first dose date as the reference start date and July 1st of birth year and truncated to a whole-year (integer) age. Patients whose derived age is less than 18 will have the required minimum age of 18 at informed confirmed; reporting for age, age groups, and lab ranges, however will be based on their derived age.
- Age group (<65, ≥65 years old)
- Age group (<65, ≥65 to <75, ≥75 to <85, ≥85 years old)
- Body Mass Index (BMI) (kg/m^2) = Weight (kg)/((Height (cm)/100)^2)
- BMI category (<25 kg/m^2, ≥25 to <30 kg/m^2, ≥30 kg/m^2)
- The duration of AD from diagnosis (years) = [(Date of informed consent – Date of AD diagnosis)+1]/ 365.25.
    - If year of onset is missing, duration of AD will be set as missing. Otherwise, unknown month will be taken as January, and unknown day will be taken as 01. The duration of AD will be rounded to 1 decimal place.
- Duration of AD (years) category (0 to <2 years, 2 to < 5 years, 5 to <10 years, 10 to <20 years, ≥20 years)
- Diagnosis age (years), derived using diagnosis date as the reference start date and July 1 of birth year and truncated to a whole-integer age.
- Diagnosis age group (<18, ≥18 to <50, ≥50 years old)
- Change from baseline = postbaseline measurement at Visit x – baseline measurement.
    - If a baseline value is missing, it will not be imputed and the change from baseline will not be calculated.
- Percent change from baseline at Visit x:
  ((Post-baseline measurement at Visit x - Baseline measurement)/Baseline measurement)*100.
    - If a baseline value is missing, it will not be imputed and percent change from baseline will not be calculated.
- Weight (kg) = weight (lbs) * 0.454.
- Weight category (<60 kg, ≥60 to <100 kg, ≥100 kg)
- Height (cm) = height (in) * 2.54.
- Cyclosporine inadequate efficacy (yes, no)
    - Set **yes** if the reason for discontinuation is inadequate response.
- Cyclosporine intolerance (yes, no)
    - Set **yes** if the reasons for discontinuation are: intolerance to medication or contraindication (Physician indicated cyclosporine was used and a contraindication was noted).
- Cyclosporine contraindication [ineligible] (yes, no)

    - Set to **yes** if cyclosporine never used because of a contraindication
- Cyclosporine inadvisable (yes, no)
  - Set to **yes** if the following reasons were selected for either not using the medication or discontinuing the medication:
    - Reason for not using medication: Physician decision, concern about side effects, unfavorable benefit risk, contraindication.
    - Reasons for discontinuation: inadequate response, intolerance to medication, or contraindication.
- TCNI inadequate efficacy (yes, no)
    - Set **yes** if the reason for discontinuation is inadequate response.
- TCNI intolerance (yes, no)
    - Set **yes** if the reasons for discontinuation are: intolerance to medication or contraindication (Physician indicated TCNI was used and a contraindication was noted).
- TCNI contraindication / [ineligible](yes, no)
    - Set to **yes** if TCNI never used because of a contraindication
- TCNI inadvisable (yes, no)
  - Set to **yes** if the following reasons were selected for either not using the medication or discontinuing the medication:
    - Reason for not using medication: Physician decision, concern about side effects, unfavorable benefit risk, contraindication.
    - Reasons for discontinuation: inadequate response, intolerance to medication, or contraindication.

6.3. Covariate Adjustment

The randomization to treatment groups at Week 0 (Visit 2) is stratified by disease severity (IGA) and geographic region as described in Section 5.1. Unless otherwise specified, the statistical analysis models will adjust for these stratification variables. The covariates used in the logistic model for categorical data will include the parameter value at baseline. The covariates used in the ANCOVA model for continuous data will include the parameter value at baseline. Inclusion of baseline in the model ensures treatment LSM are estimated at the same baseline value. When an MMRM analysis is performed, baseline value and baseline-by-visit interactions will be included as covariates.

6.4. Handling of Dropouts or Missing Data

Intercurrent events (ICH E9 R1) are events which occur after the treatment initiation and make it impossible to measure a variable or influence how it would be interpreted.

Depending on the estimand being addressed, different methods will be used to handle missing data as a result of intercurrent events. Intercurrent events can occur through the following:

- application of one of the censoring rules (including after permanent study drug discontinuation or after rescue therapy)
- discontinuation

- missing an intermediate visit prior to discontinuation or rescue
- lost to follow-up.

Non-censor intercurrent events are events that are not due to the application of any censoring rule, i.e., the last 3 items in the list above.

Note that as efficacy and health outcome data can accrue after a patient permanently discontinues study drug or begins rescue therapy, specific general censoring rules to the data will be applied to all efficacy and health outcome observations subsequent to these events depending on the estimand being addressed. These specific censoring rules are described below.

The *primary censoring rule* will censor efficacy and health outcome data after permanent study drug discontinuation or after rescue therapy. This censoring rule will be applied to all continuous and categorical efficacy and health outcome endpoints. This censoring rule is equivalent to using all the data up to rescue.

A *secondary censoring rule* will only censor efficacy and health outcome data after permanent study drug discontinuation. This sensitivity analysis will include all observed values up to study drug discontinuation. The secondary censoring rule will be applied to primary and key secondary efficacy and health outcome endpoints as sensitivity analyses.

Table JAIY.6.1 describes the planned imputation methods for efficacy and health outcome endpoints with associated censoring rules. Sections 6.4.1 through 6.4.5 summarize the methodology of each imputation rule.

Table JAIY.6.1. Imputation Techniques for Various Variables

Efficacy and Health Outcome Endpoints	Imputation Method
IGA(0,1), EASI75, 4-point Itch NRS improvement, EASI90, SCORAD75	NRI[ab], pMI[a], Tipping point[a]
EASI percent change, ADSS Item 2 change, Skin Pain NRS change	MMRM[ab], mLOCF[a], pMI[a]
All remaining categorical measures	NRI[a]
All remaining continuous efficacy and health outcome measures	MMRM[a], mLOCF[a]

Abbreviations: ADSS = Atopic Dermatitis Sleep Scale; EASI = Eczema Area and Severity Index score; IGA = Investigator's Global Assessment for AD; mLOCF = modified last observation carried forward; MMRM = mixed model repeated measures; NRI = nonresponder imputation, NRS = Numeric Rating Scale; pMI = placebo multiple imputation; SCORAD = SCORing Atopic Dermatitis.

a Analyses utilizing the primary censoring rule.

b Analyses utilizing the secondary censoring rule.

6.4.1. Nonresponder Imputation

A nonresponder imputation (NRI) method imputes missing values as non-responses and can be justified based on the composite strategy for handling intercurrent events (ICH E9 R1). This imputation procedure assumes the effects of treatments disappear after the occurrence of an intercurrent event defined by the associated censoring rule.

All categorical endpoints will utilize the NRI method after applying the primary censoring rule to patients who permanently discontinued study drug or were rescued (described in Section 6.4). Additionally, all primary and key secondary categorical endpoints will utilize NRI after applying the secondary censoring rule as sensitivity analyses. For analyses which utilize either of the censoring methods, randomized patients without at least 1 post-baseline observation will be defined as nonresponders for all visits.

6.4.2. Mixed Model for Repeated Measures

Mixed Model for Repeated Measures analyses will be performed on continuous endpoints to mitigate the impact of missing data. This approach assumes missing observations are missing-at-random (missingness is related to observed data) and borrows information from patients in the same treatment arm taking into account both the missingness of data through the correlation of the repeated measurements.

Essentially MMRM estimates the treatment effects had all patients remained on their initial treatment throughout the study. For this reason, the MMRM implies a different estimand (hypothetical strategy [ICH E9 R1]) than the one used for NRI on categorical outcomes.

All continuous endpoints will utilize MMRM after applying the primary censoring rule. As sensitivity analyses, all secondary continuous endpoints will also utilize MMRM after applying the secondary censoring rule (Table JAIY.6.1).

6.4.3. Modified Last Observation Carried Forward

For continuous measure, a modified last observation carried forward (mLOCF) imputation technique replaces missing data with the most recent non-missing post-baseline assessment. The specific modification to the LOCF is data after an intercurrent event will not be carried forward thus the mLOCF is applied after the specified censoring rule is implemented. The mLOCF assumes the effect of treatment remain the same after the event that caused missing data as it was just prior to the missing data event. Analyses using mLOCF require a nonmissing baseline and at least 1 postbaseline measure otherwise the data is missing for analyses purposes. Analyses using mLOCF help ensure the number of randomized patients who were assessed post-baseline is maximized and is reasonable for this data as data directly prior to an intercurrent event (such as initiation of rescue therapy or drop out) is likely a non-efficacious response.

All continuous efficacy and health outcomes endpoints will use with mLOCF imputation methodology with an ANCOVA as sensitivity analyses to the MMRM analyses.

6.4.4. Placebo Multiple Imputation

The Placebo Multiple Imputation (pMI) methodology will be used as a sensitivity analysis for the analysis of the primary efficacy endpoint (IGA 0 or 1 at Week 16) as well as the key secondary endpoints at Week 16. In these sensitivity analyses the primary censoring rule will be applied.

The pMI assumes that the statistical behavior of drug- and placebo- treated patients after the occurrence of intercurrent events will be the same as if patients were treated with placebo. Thus,

in the effectiveness context, pMI assumes no pharmacological benefit of the drug after the occurrence of intercurrent events but is a more conservative approach than mLOCF because it accounts for uncertainty of imputation, and therefore does not underestimate standard errors, and it limits bias. In the efficacy context pMI is a specific form of a missing not at random analysis and expected to yield a conservative estimate of efficacy.

In the pMI analysis, multiple imputations are used to replace missing outcomes for drug- and placebo-treated patients who have an intercurrent event using multiple draws from the posterior predictive distribution estimated from the placebo arm. The binary outcomes will then be derived from the imputed data.

Data are processed sequentially by repeatedly calling SAS® PROC MI to impute missing outcomes at visits t=1,.., T.

1. *Initialization:* Set t=0 (baseline visit)
2. *Iteration:* Set t=t+1. Create a data set combining records from drug- and placebo-treated patients with columns for covariates **X** and outcomes at visits 1,..,t with outcomes for all drug-treated patients set to missing at visit t and set to observed or imputed values at visits 1,..,t-1.
3. *Imputation:* Run Bayesian regression in SAS® PROC MI on this data to impute missing values for visit t using previous outcomes for visits 1 to t-1 and baseline covariates. Note that only placebo data will be used to estimate the imputation model since no outcome is available for drug-treated patients at visit t.
4. Replace imputed data for all drug-treated patients at visit t with their observed values, whenever available up to permanent study drug discontinuation and/or rescue (if censoring on rescue). If t < T then go to Step 2, otherwise proceed to Step 5.
5. Repeat steps 1-4, m times with different seed values to create m imputed complete data sets.

Analysis: For continuous endpoints, fit its treatment response model (MMRM) for each completed data set. For the primary and secondary key efficacy endpoints [IGA (0,1), EASI75, EASI90, SCORAD75, and 4-point improvement from baseline in Itch NRS], the binary outcomes will be derived from the imputed data for each patient before fitting the logistic regression model.

The number of imputed data sets will be m=100 and a 6-digit seed value will be pre-specified for each analysis. Within the program, the seed will be used to generate the m seeds needed for imputation. The initial seed values are given in Table JAIY.6.2.

Table JAIY.6.2. Seed Values for Multiple Imputation

Analysis	Seed value
Proportion of patients achieving IGA of 0 or 1 with a ≥2-point improvement from baseline at Week 16 using the primary censoring rule	123450
Percent change from baseline in EASI score at 16 weeks using the primary censoring rule. EASI75 and EASI90 will leverage imputation from EASI and therefore do not need a new seed number.	123451
Proportion of patients achieving SCORAD75 at 16 week using the primary censoring rule, with data up to rescue	123452
Proportions of patients achieving a 4-point improvement from baseline in Itch NRS at Week 16 using the primary censoring rule	123453
Mean change from baseline in Skin Pain NRS at Week 16 using the primary censoring rule	123454
Mean change from baseline in the score of Item 2 of the ADSS at Week 16 using the primary censoring rule	123455

Abbreviations: ADSS = Atopic Dermatitis Sleep Scale; EASI = Eczema Area and Severity Index score; IGA = Investigator's Global Assessment for AD; NRS = Numeric Rating Scale; SCORAD = SCORing Atopic Dermatitis.

The final inference on treatment difference is conducted from the multiple datasets using Rubin's combining rules, as implemented in SAS® PROC MIANALYZE.

6.4.5. Tipping Point Analyses

To investigate the missing data mechanism, sensitivity analyses using multiple imputation (MI) under the missing not at random assumption will be provided for the following primary and key secondary objectives as given in Table JAIY.6.3.

All patients in the ITT population will be included. Data after the occurrence of intercurrent events (after application of the primary censoring rule) will be set to missing.

Within each analysis, a most extreme case will be considered, in which all missing data for patients randomized to baricitinib 2 mg or 4 mg will be imputed using the worst possible result and all missing data for patients randomized to placebo will be imputed with the best possible result. Treatment differences will be analyzed using logistic regression or ANCOVA (Section 6.1) as appropriate.

For continuous variables, the following process will be used to determine the tipping point:

1. To handle intermittent missing visit data, a Markov chain Monte Carlo method (SAS® Proc MI with MCMC option) will be used to create a monotone missing pattern.
2. A set of Bayesian regressions (using SAS® Proc MI with MONOTONE option) will be used for the imputation of monotone dropouts. Starting from the first visit with at least 1 missing value, the regression models will be fit sequentially with treatment as a fixed effect and values from the previous visits as covariates.

3. A delta score is added to all imputed scores at the primary time point for patients in the baricitinib treatment groups, thus worsening the imputed value. The delta score is capped for patients based on the range of the outcome measure being analyzed.
4. Treatment differences between baricitinib and placebo are analyzed for each imputed dataset using ANCOVA (Section 6.1). Results across the imputed datasets are aggregated using SAS® Proc MIANALYZE in order to compute a p-value for the treatment comparisons for the given delta value.
5. Steps 3 and 4 are repeated, and the delta value added to the imputed baricitinib scores is gradually increased. The tipping point is identified as the delta value which leads to a loss of statistical significance (aggregated p-value >0.05) when evaluating baricitinib relative to the placebo group.

As a reference, for each delta value used in Steps 3 through 5, a fixed selection of delta values (ranging from slightly negative to slightly positive) will be added to imputed values in the placebo group, and Step 4 will be performed for the combination. This will result in a 2-d table, with the columns representing the delta values added to the imputed placebo responses, and the rows representing the delta values added to the imputed baricitinib responses. Separate 2-d tables will compare each baricitinib dose group to placebo.

A similar process will be used for the categorical variables:

1. Missing responses in the baricitinib groups will be imputed with a range of low response probabilities, including probabilities of 0, 0.1, and 0.2.
2. For missing responses in the placebo group, a range of responses probabilities (for example, probability = 0, 0.2 … 1) will be used to impute the missing values. Multiple imputed datasets will be generated for each response probability.
3. Treatment differences between baricitinib and placebo are analyzed for each imputed dataset using logistic regression (Section 6.1). Results across the imputed datasets are aggregated using SAS® Proc MIANALYZE in order to compute a p-value for the treatment comparisons for the given response probability. If the probability values do not allow for any variation between the multiple imputed datasets (for example, all missing responses in the placebo and baricitinib groups are imputed as responders and nonresponders, respectively), then the p-value from the single imputed dataset will be used.

The tipping point is identified as the response probability value within the placebo group that leads to a loss of statistical significance when evaluating baricitinib relative to placebo.

For tipping point analyses the number of imputed data sets will be m=100 and the seed values to start the pseudorandom number generator of SAS Proc MI (same values for MCMC option and for MONOTONE option) are given in Table JAIY.6.3.

Table JAIY.6.3. Seed Values for Imputation

Analysis	Seed value
Proportion of patients achieving IGA (0,1) with ≥2-point improvement at Week 16; primary censoring rule	123470
Proportion of patients achieving EASI75 at Week 16; primary censoring rule Proportion of patients achieving EASI90 at Week 16; primary censoring rule	123471
Proportions of patients achieving a 4-point improvement from baseline in Itch NRS at Week 16, primary censoring rule	123472
Proportion of patients achieving SCORAD75 at Week 16; primary censoring rule	123473

6.5. Multicenter Studies

This study will be conducted by multiple investigators at multiple sites internationally. The countries will be categorized into geographic regions, as described in Section 5.2.

For the analysis of the primary endpoint, treatment-by-region interaction will be added to the logistic regression model as a sensitivity analysis and results from this model will be compared to the primary model (without the interaction effect). If the treatment-by-region interaction is significant at a 2-sided α level of 0.1, the nature of this interaction will be inspected as to whether it is quantitative (i.e., the treatment effect is consistent in direction across all regions but not in size of treatment effect) or qualitative (the treatment is beneficial in some but not all regions). If the treatment-by-region interaction effect is found to be quantitative, results from the primary model will be presented. If the treatment-by-region interaction effect is found to be qualitative, further inspection will be used to identify in which regions baricitinib is found to be more beneficial.

6.6. Multiple Comparisons/Multiplicity

The primary and key secondary endpoints will be adjusted for multiplicity in order to control the overall family-wise Type I error rate at a 2-sided alpha level of 0.05.

The following is a list of primary and key secondary endpoints to be tested.

Primary Null Hypotheses:

- Null Hypotheses[IGA0-1]: Proportion of baricitinib 4-mg patients achieving IGA of 0 or 1 with a ≥2-point improvement from baseline at Week 16 is equal to the proportion of placebo patients achieving IGA of 0 or 1 with a ≥2-point improvement from baseline at Week 16
- Null Hypotheses[IGA0-1]: Proportion of baricitinib 2-mg patients achieving IGA of 0 or 1 with a ≥2-point improvement from baseline at Week 16 is equal to the proportion of placebo patients achieving IGA of 0 or 1 with a ≥2-point improvement from baseline at Week 16

Key Secondary Null Hypotheses:

- Null Hypotheses[EASI75]: Proportion of baricitinib 4-mg patients achieving EASI75 is equal to the proportion of placebo patients achieving EASI75 at Week 16
- Null Hypotheses[ITCH W16]: Proportion of baricitinib 4-mg patients achieving a 4-point improvement in Itch NRS is equal to the proportion of placebo patients achieving a 4-point improvement in Itch NRS at Week 16 among patients with baseline Itch NRS score ≥4
- Null Hypotheses[EASI PCFB]: Percent change from baseline in EASI score for baricitinib 4-mg patients is equal to the percent change from baseline in EASI score for placebo patients at Week 16
- Null Hypotheses[ITCH W4]: Proportion of baricitinib 4-mg patients achieving a 4-point improvement in Itch NRS is equal to the proportion of placebo patients achieving a 4-point improvement in Itch NRS at Week 4 among patients with baseline Itch NRS score ≥4
- Null Hypotheses[SCORAD75]: Proportion of baricitinib 4-mg patients achieving SCORAD75 is equal to the proportion of placebo patients achieving SCORAD75 at Week 16
- Null Hypotheses[EASI 90]: Proportion of baricitinib 4-mg patients achieving EASI90 is equal to the proportion of placebo patients achieving EASI90 at Week 16
- Null Hypotheses[PAIN NRS]: Mean change from baseline in Skin Pain NRS for baricitinib 4-mg patients is equal to the mean change from baseline in Skin Pain NRS for placebo patients at Week 16
- Null Hypotheses[ADSS2 W16]: Mean change from baseline in the score of Item 2 of the ADSS for baricitinib 4-mg patients equal to the mean change from baseline in the score of Item 2 of the ADSS for placebo patients at Week 16
- Null Hypotheses[ITCH W2]: Proportion of baricitinib 4-mg patients achieving a 4-point improvement in Itch NRS is equal to the proportion of placebo patients achieving a 4-point improvement in Itch NRS at Week 2 among patients with baseline Itch NRS score ≥4
- Null Hypotheses[ADSS2 W1]: Mean change from baseline in the score of Item 2 of the ADSS for baricitinib 4-mg patients is equal to the mean change from baseline in the score of Item 2 of the ADSS for placebo patients at Week 1
- Null Hypotheses[ITCH W1]: Proportion of baricitinib 4-mg patients achieving a 4-point improvement in Itch NRS is equal to the proportion of placebo patients achieving a 4-point improvement in Itch NRS at Week 1 among patients with baseline Itch NRS score ≥4
- Null Hypotheses[ITCH D2]: Proportion of baricitinib 4-mg patients achieving a 4-point improvement in Itch NRS is equal to the proportion of placebo patients achieving a 4-point improvement in Itch NRS at Day 2 among patients with baseline Itch NRS score ≥4
- Null Hypotheses[EASI75]: Proportion of baricitinib 2-mg patients achieving EASI75 is equal to the proportion of placebo patients achieving EASI75 at Week 16
- Null Hypotheses[ITCH W16]: Proportion of baricitinib 2-mg patients achieving a 4-point improvement in Itch NRS is equal to the proportion of placebo patients achieving a 4-point improvement in Itch NRS at Week 16 among patients with baseline Itch NRS score ≥4

- Null Hypotheses[EASI PCFB]: Percent change from baseline in EASI score for baricitinib 2-mg patients is equal to the percent change from baseline in EASI score for placebo patients at Week 16
- Null Hypotheses[ITCH W4]: Proportion of baricitinib 2-mg patients achieving a 4-point improvement in Itch NRS is equal to the proportion of placebo patients achieving a 4-point improvement in Itch NRS at Week 4 among patients with baseline Itch NRS score ≥4
- Null Hypotheses[SCORAD75]: Proportion of baricitinib 2-mg patients achieving SCORAD75 is equal to the proportion of placebo patients achieving SCORAD75 at Week 16
- Null Hypotheses[EASI90]: Proportion of baricitinib 2-mg patients achieving EASI90 is equal to the proportion of placebo patients achieving EASI90 at Week 16
- Null Hypotheses[PAIN NRS]: Mean change from baseline in Skin Pain NRS for baricitinib 2-mg patients is equal to the mean change from baseline in Skin Pain NRS for placebo patients at Week 16
- Null Hypotheses[ADSS2 W16]: Mean change from baseline in the score of Item 2 of the ADSS for baricitinib 2-mg patients is equal to the mean change from baseline in the score of Item 2 of the ADSS for placebo patients at Week 16
- Null Hypotheses [ITCH W2]: Proportion of baricitinib 2-mg patients achieving a 4-point improvement in Itch NRS is equal to the proportion of placebo patients achieving a 4-point improvement in Itch NRS at Week 2 among patients with baseline Itch NRS score ≥4
- Null Hypotheses[ADSS2 W1]: Mean change from baseline in the score of Item 2 of the ADSS for baricitinib 2-mg patients is equal to the mean change from baseline in the score of Item 2 of the ADSS for placebo patients at Week 1
- Null Hypotheses[ITCH W1]: Proportion of baricitinib 2-mg patients achieving a 4-point improvement in Itch NRS is equal to the proportion of placebo patients achieving a 4-point improvement in Itch NRS at Week 1 among patients with baseline Itch NRS score ≥4
- Null Hypotheses[ITCH D2]: Proportion of baricitinib 2-mg patients achieving a 4-point improvement in Itch NRS is equal to the proportion of placebo patients achieving a 4-point improvement in Itch NRS at Day 2 among patients with baseline Itch NRS score ≥4

The multiple testing strategy for the primary and the major secondary endpoint will be implemented through the graphical testing procedure depicted by Figure JAIY.6.1.

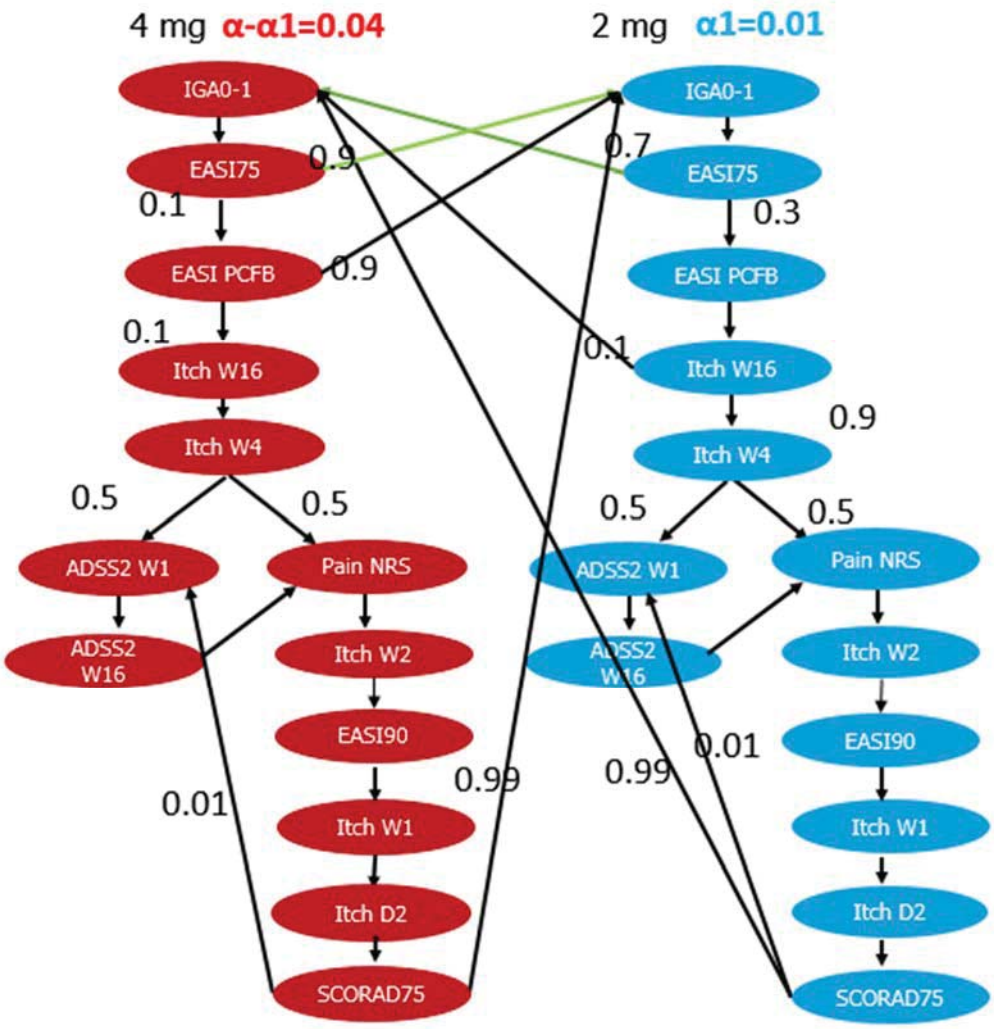


Figure JAIY.6.1. Illustration of graphical multiple testing procedure with initial α allocation and weights.

For Japan, the same testing strategy will be used, however, the proportion of patients achieving EASI75 and IGA of 0 or 1 at Week 16 are considered co-primary. Both endpoints have to meet statistical significance compared with placebo in order to demonstrate a superiority of a given baricitinib dose (baricitinib 4-mg and/or baricitinib 2-mg) compared with placebo.

There will be no adjustment for multiple comparisons for any other analyses.

6.7. Patient Disposition

An overview of patient populations will be summarized by treatment group. Frequency counts and percentages of patients excluded prior to randomization by primary reason for exclusion will be provided for patients who failed to meet study entry requirements during screening.

Patient disposition through Week 16 will be summarized using the ITT population. Frequency counts and percentages of patients who complete the study treatment visits or discontinue early from the study along with whether they completed follow-up, did not complete follow-up or

enrolled into the extension will be summarized separately by treatment group for patients who are not rescued and for patients who are rescued, along with their reason for study discontinuation. Frequency counts and percentages of patients who complete the treatment or discontinue treatment early will also be summarized separately by treatment group for patients who are not rescued and for patients who are rescued, along with their reason for treatment discontinuation.

A listing of patient disposition will be provided for all randomized patients, with the extent of their participation in the study and the reason for discontinuation. A listing of all randomized patients with their treatment assignment will also be provided.

6.8. Patient Characteristics

Patient characteristics including demographics and baseline characteristics will be summarized descriptively by treatment group for the ITT population. Historical illnesses and pre-existing conditions will be summarized descriptively by treatment group for the ITT population. No formal statistical comparisons will be made among treatment groups unless otherwise stated.

6.8.1. Demographics

Patient demographics will be summarized as described above. The following demographic information will be included:

- Age
- Age group (<65 vs. ≥65)
- Age group (<65, ≥65 to <75, ≥75 to <85, ≥85)
- Gender (male, female)
- Race (American Indian or Alaska Native, Asian, Black or African American, Native Hawaiian or Other Pacific Islander, White, Multiple)
- Region (as defined in Table JAIY.5.1)
- Country
- Weight (kg)
- Weight category (<60 kg, ≥60 to <100 kg, ≥100 kg)
- Height (cm)
- BMI (kg/m^2)
- BMI category (<25 kg/m^2, ≥25 to <30 kg/m^2, ≥30 kg/m^2)

A listing of patient demographics will also be provided for the ITT population.

6.8.2. Baseline Disease Characteristics

The following baseline disease information will be categorized and presented for baseline AD clinical characteristics, baseline health outcome measures, and other baseline demographic and disease characteristics as described above:

- Duration since AD diagnosis (years)
- Duration since AD diagnosis category (0 to <2 years, 2 to <5 years, 5 to <10 years, 10 to <20 years, ≥20 years)

- Age at Diagnosis (years)
- Age Group at Diagnosis (<18 years, ≥18 to <50 years, ≥50 years)
- Habits (Alcohol: Never, Current, Former; Tobacco: Never, Current, Former)
- Skin Infections treated with a pharmacological agent within past year (yes, no, unknown; number if yes)
- Atopic Dermatitis Flares within past year (yes, no, unknown; number if yes)
- Validated Investigator's Global Assessment for AD (IGA) score
- Eczema Area and Severity Index (EASI) score
- SCORing Atopic Dermatitis (SCORAD)
- Body Surface Area (BSA) affected by AD
- Hospital Anxiety Depression Scale (HADS) subscales
- Patient-Oriented Eczema Measure (POEM)
- Itch Numerical Rating Scale (NRS)
- Atopic Dermatitis Sleep Scale (ADSS) Item 2
- Dermatology Life Quality Index (DLQI)
- Skin Pain NRS
- Patient Global Impression of Severity (PGI-S-AD)
- Prior therapy (topical therapy only; systemic therapy)
- Prior use of Cyclosporine (yes, no)
- Cyclosporine inadequate response (yes, no)
- Cyclosporine intolerance (yes, no)
- Cyclosporine contraindication [ineligible] (yes, no)
- Cyclosporine inadvisable (yes no)
- Prior use of TCNI (yes, no)
- TCNI inadequate response (yes, no)
- TCNI intolerance (yes, no)
- TCNI contraindication [ineligible] (yes, no)
- TCNI inadvisable (yes, no)
- Vaccine (yes, no)
- Baseline renal function status: impaired (eGFR <60 mL/min/1.73 m^2) or not impaired (eGFR ≥60 mL/min/1.73 m^2)
- Immunoglobulin E (IgE): intrinsic(<200 kU/I) or extrinsic (≥200 kU/I)

6.8.3. Historical Illness and Pre-existing Conditions

Historical illnesses are defined as those conditions recorded in the Pre-existing Conditions and Medical History electronic case report form (eCRF) or from the Prespecified Medical History: Comorbidities eCRF with an end date prior to the informed consent date. The number and percentage of patients with selected historical diagnoses will be summarized by treatment group using the ITT population. Historical diagnoses will be categorized using the Medical Dictionary for Regulatory Activities (MedDRA®, most current available version) algorithmic standardized MedDRA queries (SMQs) or similar pre-defined lists of preferred terms (PTs) of interest.

Preexisting conditions are defined as those conditions with a start date prior to the first dose of the study drug and stop dates that are at or after the informed consent date or have no stop date

(i.e., are ongoing). For events occurring on the day of the first dose of study treatment, the date and time of the onset of the event will both be used to determine if the event was pre-existing. Conditions with a partial or missing start date (or time if needed) will be assumed to be 'not pre-existing' unless there is evidence, through comparison of partial dates, to suggest otherwise. Pre-existing conditions will be categorized using the MedDRA SMQs or similar pre-defined lists of PTs of interest. Frequency counts and percentages of patients with selected pre-existing conditions will be summarized by treatment group using the ITT population.

6.9. Treatment Compliance

Patient compliance with study medication will be assessed from Week 0 (Visit 2) to Week 16 (Visit 8) or Early Termination using the ITT population.

All patients are expected to take 2 tablets daily from a package as described in the protocol. Each bottle contains 36 tablets. A patient is considered noncompliant if he or she misses >20% of the prescribed doses during the study, unless the patient's study drug is withheld by the investigator. For patients who had their treatment temporarily interrupted by the investigator, the period of time that dose was withheld will be taken into account in the compliance calculation.

Compliance in the period of interest up to Visit *x* will be calculated as follows:

$$\text{Compliance} = \frac{\text{total number of tablets dispensed} - \text{total number of tablets returned}}{\text{expected number of total tablets}}$$

where

- Total number of tablets dispensed: sum of tablets dispensed in the period of interest prior to Visit *x*;
- Total number of tablets returned: sum of the tablets returned in the period of interest prior to and including Visit *x*;
- Expected number of tablets: number of days in the period of interest*number of tablets taken per day = [(date of last dose – date of first dose + 1) – number of days of temporary drug interruption]*number of tablets taken per day

Patients who are significantly noncompliant (compliance <80%) through Week 16 will be excluded from the PPS population.

Descriptive statistics for percent compliance and non-compliance rate will be summarized for the ITT population by treatment group for Week 0 through Week 16. Sub-intervals of interest, such as compliance between visits, may also be presented. The number of expected doses, tablets dispensed, tablets returned, and percent compliance will be listed by patient for Week 0 through Week 16.

6.10. Rescue Therapy

Rescue therapy with additional topical and systemic therapies is available starting after 2 weeks of treatment (Visit 4), for patients who are experiencing worsening and unacceptable symptoms of AD despite treatment with IP and moderate-potency TCS. Patients whose lesions persist or worsen despite the use of emollients and background TCS therapy (e.g., triamcinolone 0.1% cream and/or hydrocortisone 2.5% ointment) and/or patients who require prolonged applications

of triamcinolone 0.1% cream (moderate-potency TCS) on large surfaces may be considered for rescue to high- or ultra-high-potency TCS. If topical rescue therapy as described above fails to sufficiently control AD symptoms, then oral systemic medications may be used as rescue (e.g., corticosteroids, cyclosporine, methotrexate); however, investigational product will be required to be permanently discontinued for the remainder of the 16-week study duration. If these medications are needed for other medical conditions (e.g., asthma flare), they will still be treated as rescue medications.

The initial rescue therapy will be the first non-missing record before the last dose date from the CRF page *Concomitant Therapy: Rescue Therapy*.

A summary of the initial rescue therapy and the reason for requiring initial rescue will be produced, as well as a summary of the proportion of patients initially rescued at each study visit. A summary of all rescue medications will be provided.

6.11. Previous and Concomitant Therapy

Summaries of previous and concomitant medications will be based on the ITT population.

At screening, previous and current AD treatments are recorded for each patient. Concomitant therapy for the treatment period is defined as therapy that starts before or during the treatment period and ends during the treatment period or is ongoing (has no end date or ends after the treatment period). Should there be insufficient data to make this comparison (for example, the concomitant therapy stop year is the same as the treatment start year, but the concomitant therapy stop month and day are missing), the medication will be considered as concomitant for the treatment period.

Summaries of previous medications will be as follows:

- Previous AD therapies

Summaries of concomitant medications, with sponsor and non-sponsor provided background TCS included, will be as follows:

- General Concomitant medications excluding rescue medicine

6.11.1. Background TCS

Background TCS therapy with moderate-potency and/or low-potency TCS (e.g., triamcinolone 0.1% cream and/or hydrocortisone 2.5% ointment) are to be used on active lesions, as described in Section 7.7.3 in the protocol.

The dispensed weight of sponsor-provided TCS tubes for the two different potencies (low and moderate) varies between countries due to different supply regions. Average weights of full tubes were used to determine the dispensed weights for each region. Returned tubes were weighed with cap (without the carton) to determine the amount of TCS in grams (g) used at each visit.

For low potency TCS, the dispensed tube weight with cap (without the carton) in Japan is 13.5g. For countries supplied by European distributors (Austria, Germany, Italy, Poland, and Spain), the dispensed weight of low potency TCS is 21g. The remaining countries, supplied by US distributors (Argentina, Australia, Korea, and Taiwan), the weight of low potency TCS is 40g.

For moderate potency TCS, the dispensed tube weight with cap (without the carton) in Japan is 13.5g. For countries supplied by European distributors, the dispensed weight of moderate potency TCS is 38g. The remaining countries, supplied by US distributors, the weight of moderate potency TCS is 40g.The total amount of background TCS, provided by sponsor, will be summarized in grams by potency (low and moderate) and both potencies, between visits (Week 0 through Week 1, Week 1 through Week 2, Week 2 through Week 4, Week 4 through Week 8, Week 8 through 12, Week 12 through Week 16), and throughout the entire 16-week treatment period. If a returned tube is not weighed or not returned, then the tube can be classified as partially used, fully used, unused, or unknown. Partially used rescue medication tubes will be defined as 50% used whereas fully used and unused tubes will be defined as 100% and 0% used respectively. When drug accountability is not performed for a particular tube of rescue medication or an answer of 'unknown' is given for a tube which is not returned, that particular tube will not be included in the analysis. The main analysis on the total amount of background TCS throughout the entire 16-week treatment period will apply censoring rule #1. After patients who get rescued or discontinue IP, whichever is earlier, it is assumed that they would use the same amount of TCS as they did before. Analysis will be done via analysis of variance (ANOVA), with geographic region, baseline disease severity and treatment as factors in the model. The secondary analysis will apply censoring rule #2 with the same assumptions as described above.

Whether any background TCS is used or not used for each patient is also collected on the diary device in each day starting from the first dose date throughout the study.

The total number of days that the patients did not use background TCS will be summarized by both potencies throughout the entire 16-week treatment period. The main analysis applies censoring rule #1. After patients who are rescued or discontinue IP, it is assumed that background TCS would be applied each day. In case of missing values in the daily diary, it will be assumed that background TCS has been used. Analysis will be done via ANOVA, with geographic region, baseline disease severity and treatment as factors in the model. A secondary analysis will apply censoring rule #2, with the same assumptions for missing values as described above.

6.12. Efficacy Analyses

The general methods used to summarize efficacy data, including the definition of baseline value for assessments are described in Section 6.2. The censoring rules applied to data as well as imputation methods are described in Section 6.4.

Table JAIY.6.4 provides the descriptions and derivations of the primary, secondary, and exploratory efficacy outcomes.

Table JAIY.6.5 provides the detailed analyses including analysis type, method and imputation, population, time point, and comparisons for efficacy analyses.

Table JAIY.6.4. Description and Derivation of Primary, Secondary and Exploratory Efficacy Outcomes

Measure	Description	Variable	Derivation / Comment	Imputation Approach if Missing Components
Validated Investigator's Global Assessment for AD (IGA)	The validated Investigator's global assessment of the patient's overall severity of their AD, based on a static, numeric 5-point scale from 0 (clear) to 4 (severe). The score is based on an overall assessment of the degree of erythema, papulation/induration, oozing/crusting, and lichenification.	IGA score	Single item. Range: 0 to 4 0 represents "clear" 4 represents "severe"	Single item, missing if missing.
		Change from baseline in IGA score	Change from baseline: observed IGA score – baseline IGA score	Missing if baseline or observed value is missing.
		▪ IGA [0,1] with ≥2-point improvement ▪ IGA [0]	▪ Observed score of 0 or 1 and change from baseline ≤2 ▪ Observed score of 0	▪ Missing if baseline or observed value is missing. ▪ Single item, missing if missing.

Eczema Area and Severity Index (EASI)	The EASI assesses objective physician estimates of 2 dimensions of atopic dermatitis – disease extent and clinical signs (Hanifin et al 2001) – by scoring the extent of disease (percentage of skin affected: 0 = 0%; 1 = 1-9%; 2 = 10-29%; 3 = 30-49%; 4 = 50-69%; 5 = 70-89%; 6 = 90-100%) and the severity of 4 clinical signs (erythema, edema/papulation, excoriation, and lichenification) each on a scale of 0 to 3 (0 = none, absent; 1 = mild; 2 = moderate; 3 = severe) at 4 body sites (head and neck, trunk, upper limbs, and lower limbs). Half scores are allowed between severities 1, 2 and 3. Each body site will have a score that ranges from 0 to 72, and the final EASI score will be obtained by weight-averaging these 4 scores. Hence, the final EASI score will range from 0 to 72 for each time point.	EASI score	Derive EASI region score for each of head and neck, trunk, upper limbs, and lower limbs as follows: $EASI_{region}$ = (Erythema + edema/papulation + Excoriation + Lichenification) *(value from percentage involvement), where erythema, edema/papulation, excoriation, and lichenification are evaluated on a scale of 0 to 3 and value from percentage involvement is on a scale of 0 to 6. Then total EASI score is as follows: $EASI = 0.1*EASI_{head\ and\ neck} + 0.3*EASI_{trunk} + 0.2*EASI_{upper\ limbs} + 0.4*EASI_{lower\ limbs}$	N/A – partial assessments cannot be saved.
		▪ Change from baseline in EASI score ▪ Percent change from baseline EASI score	Change from baseline: observed EASI score – baseline EASI score % change from baseline: $100 \times \frac{Observed\ score - Baseline}{Baseline}$	Missing if baseline or observed value is missing.
		EASI50	% Improvement in EASI score from baseline ≥ 50%: % change from baseline ≤ -50	Missing if baseline or observed value is missing.
		EASI75	% Improvement in EASI score from baseline ≥75%: % change from baseline ≤-75	Missing if baseline or observed value is missing.
		EASI90	% Improvement in EASI score from baseline ≥90%: % change from baseline ≤-90	Missing if baseline or observed value is missing.
Body Surface Area (BSA) Affected by AD	Body surface area affected by AD will be assessed for 4 separate body regions and is collected as part of the EASI assessment: head and neck, trunk (including genital region), upper extremities, and lower extremities	BSA score	Use the percentage of skin affected for each region (0 to 100%) in EASI as follows: $BSA\ Total = 0.1*BSA_{head\ and\ neck} + 0.3*BSA_{trunk} + 0.2*BSA_{upper\ limbs} +$	N/A – partial assessments cannot be saved.

	(including the buttocks). Each body region will be assessed for disease extent ranging from 0% to 100% involvement. The overall total percentage will be reported based off of all 4 body regions combined, after applying specific multipliers to the different body regions to account for the percent of the total BSA represented by each of the 4 regions.		0.4*$BSA_{lower\ limbs}$	
		Change from baseline in BSA score	Change from baseline: observed BSA score – baseline BSA score	Missing if baseline or observed value is missing.
SCORing Atopic Dermatitis (SCORAD)	The SCORing Atopic Dermatitis (SCORAD) index uses the rule of nines to assess disease extent (head and neck 9%; upper limbs 9% each; lower limbs 18% each; anterior trunk 18%; back 18%; and genitals 1%). It evaluates 6 clinical characteristics to determine disease severity: (1) erythema, (2) edema/papulation, (3) oozing/crusts, (4) excoriation, (5) lichenification, and (6) dryness on a scale of 0 to 3 (0=absence, 1=mild, 2=moderate, 3=severe). The SCORAD index also assesses subjective symptoms of pruritus and sleep loss in the last 72 hours on visual analogue scales (VAS) of 0 to 10 where 0 is no itch or sleep loss and 10 is worst imaginable itch or sleep loss. These 3 aspects: extent of disease, disease severity, and subjective symptoms combine to give a maximum possible score of 103 (Stalder et al. 1993; Kunz et al. 1997; Schram et al. 2012).	SCORAD score	SCORAD = A/5 + 7B/2 + C, where A is extent of disease, range 0-100 B is disease severity, range 0-18 C is subjective symptoms, range 0-20	Missing if components A and B are missing or if component C is missing. Partial assessments performed by physician cannot be saved and partial assessments performed by subject cannot be saved.
		▪ Change from baseline in SCORAD score ▪ Percent change from baseline in SCORAD score	Change from baseline: observed SCORAD score – baseline SCORAD score % change from baseline: $100 \times \frac{Observed\ score - Baseline}{Baseline}$	Missing if baseline or observed value is missing.
		SCORAD75	% Improvement in SCORAD from baseline ≥75%: % change from baseline ≤-75	Missing if baseline or observed value is missing.
		SCORAD90	% Improvement in SCORAD from baseline ≥90%: % change from baseline ≤-90	Missing if baseline or observed value is missing.

Table JAIY.6.5. Description of Primary, Secondary and Exploratory Efficacy Analyses

Measure	Variable	Analysis Method (Section 6.2.3)	Population (Section 6.2.1)	Comparison/Time Point	Analysis Type
Validated Investigator's Global Assessment for AD (IGA)	Proportion of patients achieving IGA [0,1] with a ≥2-point improvement	Logistic regression using NRI	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Primary analysis
			PPS	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Sensitivity analysis
			ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 4	Secondary analysis
		Logistic regression using pMI and Tipping Point	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Sensitivity analysis
	Proportion of patients achieving IGA [0]	Logistic regression using NRI	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Secondary analysis
Eczema Area and Severity Index (EASI)	• EASI score • Change from baseline in EASI score • Percent change from baseline in EASI score	MMRM	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Key secondary analysis
			PPS	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Sensitivity analysis
		ANCOVA using mLOCF	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Sensitivity analysis
		pMI and Tipping Point	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Sensitivity analysis
	• Proportion of patients achieving EASI50	Logistic regression using NRI	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Secondary analysis
	• Proportion of patients achieving EASI75 • Proportion of patients achieving EASI90	Logistic regression using NRI	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Key secondary analysis
			PPS	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Sensitivity analysis
		pMI and Tipping Point	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Sensitivity analysis
Body Surface Area (BSA) Affected by AD	• BSA score • Change from baseline in BSA score	MMRM	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Secondary analysis
		ANCOVA using mLOCF	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Sensitivity analysis

Measure	Variable	Analysis Method (Section 6.2.3)	Population (Section 6.2.1)	Comparison/Time Point	Analysis Type
SCORing Atopic Dermatitis (SCORAD)	• SCORAD score • Change from baseline in SCORAD score • Percent change from baseline in SCORAD score	MMRM	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Secondary analysis
		ANCOVA using mLOCF	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Sensitivity analysis
	Proportion of patients achieving SCORAD75	Logistic regression using NRI	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Key secondary analysis
			PPS	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Sensitivity analysis
		Logistic regression using pMI	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Sensitivity analysis
	Proportion of patients achieving SCORAD90	Logistic regression using NRI	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Secondary analysis
		pMI and Tipping Point	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Sensitivity analysis
Skin Infections	Proportion of patients developing skin infections requiring antibiotic treatment	Fisher's exact	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Secondary analysis

Abbreviations: ANCOVA = analysis of covariance; Bari = baricitinib; ITT = intent-to-treat; mLOCF = modified last observation carried forward; MMRM = mixed model repeated measures; NRI = nonresponder imputation; PBO = placebo; pMI=placebo multiple imputation; PPS = per protocol set.

Notes: (1) for all other post-baseline visits not mentioned in the table, but collected for the measures as specified in the protocol, the analyses will be made as exploratory analyses.

(2) All primary and key secondary analyses will be performed for Japan population. Other key secondary and exploratory analysis may be performed for Japan population.

6.12.1. Primary Outcome and Methodology

The validated Investigator's Global Assessment for AD (IGA) uses the clinical characteristics of erythema, papulation/induration, oozing/crusting and lichenification to produce a single-item score ranging from 0 to 4. The primary analysis of the study is to test the null hypotheses that neither baricitinib 4 mg nor baricitinib 2 mg is superior to placebo when evaluating the proportion of patients achieving IGA of 0 or 1 at Week 16 in the ITT population. The analysis assumes that treatment response disappears after patients are rescued or permanently discontinue from treatment. This will serve as the primary estimand. In this estimand, missing data due to the application of the primary censoring rule and the occurrence of other non-censor intercurrent events will be imputed using the NRI method described in Section 6.4.1.

A supplemental estimand is to test the null hypotheses that neither baricitinib 4 mg nor baricitinib 2 mg is superior to placebo when evaluating the proportion of patients achieving IGA of 0 or 1 at Week 16 in the ITT population. This analysis assumes the treatment response disappears after patients permanently discontinue from treatment. In this supplemental estimand, missing data due to the application of the secondary censoring rule and the occurrence of other non-censor intercurrent events will be imputed using the NRI method described in Section 6.4.1.

A logistic regression analysis as described in Section 6.2.3 will be used for the comparisons. The odds ratio, the corresponding 95% CIs and p-value, as well as the treatment differences and the corresponding 95% CIs, will be reported.

Multiplicity controlled analyses will be performed on the primary and key secondary (see Section 4.2.1) objectives to control the overall Type I error rate at a 2-sided alpha level of 0.05. A graphical approach will be used to perform the multiplicity controlled analyses as described in Section 6.6.

6.12.2. Secondary and Exploratory Efficacy Analyses

For secondary analysis, the null hypotheses is that neither baricitinib 4 mg nor baricitinib 2 mg is superior to placebo in the ITT population. These analyses assume treatment response disappears after patients are rescued or permanently discontinued from treatment and will serve as the primary estimand. In this estimand, missing data due to the application of the primary censoring rule and the occurrence of other non-censor intercurrent events will be imputed using the method described in Table JAIY.6.1.

A supplemental estimand for secondary endpoints is to test the null hypotheses that neither baricitinib 4 mg nor baricitinib 2 mg is superior to placebo in the ITT population. These analyses assume the treatment response disappears after patients permanently discontinue from treatment. In this supplemental estimand, missing data due to the application of the secondary censoring rule and the occurrence of other non-censor intercurrent events will be imputed using the method described in Table JAIY.6.1.

A list of exploratory endpoints are provided in Section 4.2.2. There will be no adjustment for multiple comparisons for exploratory endpoints. The secondary and exploratory efficacy

endpoints are detailed in Table JAIY.6.4 and analyses are provided in Table JAIY.6.5. Health outcomes analyses are described in Section 6.13.

6.12.3. Sensitivity Analyses

Sensitivity analyses are included to demonstrate robustness of analyses methods using different missing data imputations, censoring rules, populations and analyses assumptions. Sensitivity analyses for select outcomes have been previously described and include the following:

- Analyses of key endpoints using the per-protocol analysis set (Section 6.2.1)
- Analyses of key endpoints using the secondary censoring rule (Section 6.2)
- Placebo multiple imputation (Section 6.4.4)
- Tipping point analysis (Section 6.4.5)
- The addition of a treatment-by-region interaction to the logistic regression model for the primary outcome (Section 6.5)
- Analysis of continuous outcomes with ANCOVA (Section 6.2.3), with missing data imputed using mLOCF (Section 6.4.3).

6.13. Health Outcomes/Quality-of-Life Analyses

The general methods used to summarize health outcomes and quality-of-life measures, including the definition of baseline value for assessments are described in Section 6.1.

Health outcomes and quality-of-life measures will generally be analyzed according to the formats discussed in Section 6.12.

Table JAIY.6.6 includes the descriptions and derivations of the health outcomes and quality-of-life measures.

Table JAIY.6.7 provides the detailed analyses including analysis type, method and imputation, population, time point, and comparisons for health outcomes and quality-of-life measures.

Table JAIY.6.6. Description and Derivation of Health Outcomes and Quality-of-Life Measures

Measure	Description	Variable	Derivation / Comment	Imputation Approach if Missing Components
Itch Numeric Rating Scale (NRS)	The Itch Numeric Rating Scale (NRS) is a patient-administered, 11-point horizontal scale anchored at 0 and 10, with 0 representing "no itch" and 10 representing "worst itch imaginable." Overall severity of a patient's itching is indicated by selecting the number that best describes the worst level of itching in the past 24 hours (Naegeli et al. 2015; Kimball et al. 2016). Refer to Section 6.2.2 for details on how to calculate the weekly score which will be used in the continuous analysis.	Itch NRS score	Single item; range 0-10. Refer to Section 6.2.2 on how to derive the visit score.	Refer to Section 6.2.2 on how to derive the weekly visit score.
		▪ Change from baseline in Itch NRS ▪ Percent change from baseline in Itch NRS	Change from baseline: observed Itch score – baseline Itch score % change from baseline: $100 \times \frac{Observed\ score - Baseline}{Baseline}$	Missing if baseline or observed value is missing.
		4-point Itch improvement in subgroup of patients with baseline Itch NRS ≥4	Change from baseline ≤-4 and baseline ≥4	Missing if baseline is missing or <4 or observed value is missing.
		Itch-free days (Itch NRS = 0)	The number of itch-free days during intervals starting on the day of the first study drug administration. This will be calculated for the following intervals: baseline to Week 4, Week 4 to Week 8, Week 8 to Week 12 and Week 12 to Week 16. Day 1 is defined as the day of first study drug administration therefore the baseline to Week 4 assessment is based on Day 1 to Day 28, Week 4 to Week 8 is based on Day 29 to Day 56, etc.	Missing if observed value is missing.

Measure	Description	Variable	Derivation / Comment	Imputation Approach if Missing Components
Skin Pain Numeric Rating Scale (NRS)	Skin Pain NRS is a patient-administered, 11-point horizontal scale anchored at 0 and 10, with 0 representing "no pain" and 10 representing "worst pain imaginable." Overall severity of a patient's skin pain is indicated by selecting the number that best describes the worst level of skin pain in the past 24 hours Refer to Section 6.2.2 for details on how to calculate the weekly score which will be used in the continuous analysis.	Skin Pain NRS score	Single item; range 0 to 10. Refer to Section 6.2.2 on how to derive the visit score.	Refer to Section 6.2.2 on how to derive the visit score.
		Change from baseline in Skin Pain NRS	Change from baseline: observed skin pain score – baseline skin pain score	Missing if baseline or observed value is missing.
		Skin Pain-free days (Skin Pain NRS = 0)	The number of skin pain-free days during intervals starting on the day of the first study drug administration. This will be calculated for the following intervals: baseline to Week 4, Week 4 to Week 8, Week 8 to Week 12 and Week 12 to Week 16. Thus, if Day 1 is defined as the day of first study drug administration, the baseline to Week 4 assessment is based on Day 1 to Day 28, Week 4 to Week 8 is based on Day 29 to Day 56, etc.	Missing if observed value is missing.

Measure	Description	Variable	Derivation / Comment	Imputation Approach if Missing Components
Atopic Dermatitis Sleep Scale (ADSS)	The Atopic Dermatitis Sleep Scale (ADSS) is a 3-item, patient-administered questionnaire developed to assess the impact of itch on sleep including difficulty falling asleep, frequency of waking, and difficulty getting back to sleep last night. Patient's rate their difficulty falling asleep and difficulty getting back to sleep, items 1 and 3, respectively, using a 5-point Likert-type scale with response options ranging from 0 "not at all" to 4 "very difficult." Patients report their frequency of waking last night, item 2, by selecting the number of times they woke up each night, ranging from 0 to 29 times. The ADSS is designed to be completed each day with respondents thinking about sleep "last night." Each item is scored individually.	▪ Item 1 score of ADSS ▪ Item 2 score of ADSS ▪ Item 3 score of ADSS	Single items: Item 1, range 0 to 4; Item 2, range 0 to 29; Item 3, range 0 to 4. Refer to Section 6.2.2 on how to derive the visit score.	Refer to Section 6.2.2 on how to derive the weekly visit score.
		▪ Change from baseline in score of Item 1 of ADSS ▪ Change from baseline in score of Item 2 of ADSS ▪ Change from baseline in score of Item 3 of ADSS	Change from baseline: observed ADSS item score – baseline ADSS item score	Missing if baseline or observed value is missing.

Measure	Description	Variable	Derivation / Comment	Imputation Approach if Missing Components
Patient- Oriented Eczema Measure (POEM)	The POEM is a simple, 7-item, patient-administered scale that assesses disease severity in children and adults. Patients respond to questions about the frequency of 7 symptoms (itching, sleep disturbance, bleeding, weeping/oozing, cracking, flaking, and dryness/roughness) over the last week. Response categories include "No days," "1-2 days," "3-4 days," "5-6 days," and "Every day" with corresponding scores of 0, 1, 2, 3, and 4, respectively. Scores range from 0-28 with higher total scores indicating greater disease severity (Charman et al. 2004).	POEM score	POEM total score: sum of questions 1 to 7, Range 0 to 28.	If a single question is left unanswered, then that question is scored as 0. If more than one question is unanswered, then the tool is not scored. If more than one response is selected, then the response with the highest score is used.
		Change from baseline in POEM score	Change from baseline: observed POEM score – baseline POEM score	Missing if baseline or observed value is missing.
		4-point improvement in POEM score in subgroup of patients with baseline ≥4	Change from baseline ≤-4 and baseline ≥4	Missing if baseline is missing or <4 or observed value is missing.

Measure	Description	Variable	Derivation / Comment	Imputation Approach if Missing Components
Patient Global Impression of Severity–Atopic Dermatitis (PGI-S-AD)	The Patient Global Impression of Severity–Atopic Dermatitis (PGI-S-AD) is a single-item question asking the patient how they would rate their overall AD symptoms over the past 24 hours. The 5 categories of responses range from "no symptoms" to "severe."	PGI-S-AD score	Single item. Range 1 to 5. Refer to Section 6.2.2 on how to derive the visit score.	Refer to Section 6.2.2 on how to derive the visit score.
		Change from baseline in PGI-S-AD	Change from baseline: observed PGI-S-AD score – baseline PGI-S-AD score	Missing if baseline or observed value is missing.
Hospital Anxiety Depression Scale (HADS)	The Hospital Anxiety Depression Scale (HADS) is a 14-item self-assessment scale that determines the levels of anxiety and depression that a patient is experiencing over the past week. The HADS utilizes a 4-point Likert scale (e.g., 0 to 3) for each question and is intended for ages 12 to 65 years (Zigmond and Snaith 1983; White et al. 1999). Scores for each domain (anxiety and depression) can range from 0 to 21, with higher scores indicating greater anxiety or depression (Zigmond and Snaith 1983; Snaith 2003).	HADS score for anxiety and depression domains	Anxiety domain score is sum of the seven anxiety questions, range 0 to 21; Depression domain score is sum of the seven depression questions, range 0 to 21.	N/A – partial assessments cannot be saved.
		Change from baseline in HADS total score, anxiety and depression domain	Change from baseline: observed HADS domain score – baseline HADS domain score	Missing if baseline or observed value is missing.
Dermatology Life Quality Index (DLQI)	The Dermatology Life Quality Index (DLQI) is a simple, patient-administered, 10-item, validated, quality-of-life questionnaire that covers 6 domains including symptoms and feelings, daily activities, leisure, work and school, personal relationships, and treatment. The recall period of this scale is over the "last week." Response categories include "a little," "a lot," and "very much," with corresponding	Symptoms and feelings domain	Sum of questions 1 and 2, range 0 to 6.	N/A – partial assessments cannot be saved.
		Daily activities domain	Sum of questions 3 and 4, range 0 to 6.	N/A – partial assessments cannot be saved.
		Leisure domain	Sum of questions 5 and 6, range 0 to 6.	N/A – partial assessments cannot be saved.

Measure	Description	Variable	Derivation / Comment	Imputation Approach if Missing Components
	scores of 1, 2, and 3, respectively, and "not at all," or unanswered ("not relevant") responses scored as 0. Scores range from 0-30 with higher scores indicating greater impairment of quality of life. A DLQI total score of 0 to 1 is considered as having no effect on a patient's health-related QoL (Hongbo et al. 2005), and a 4-point change from baseline is considered as the minimal clinically important difference threshold (Khilji et al. 2002; Basra et al. 2015).	Work and school domain	Sum of questions 7 and 7B (if it is answered), range 0 to 3. Responses of "yes" and "no" on Question 7 are given scores of 3 and 0 respectively. If Question 7 is answered "no" then Question 7b is answered with "a lot", "a little", "not at all" getting scores of 2, 1, 0 respectively.	N/A – partial assessments cannot be saved.
		Personal relationships domain	Sum of questions 8 and 9, range 0 to 6.	N/A – partial assessments cannot be saved.
		Treatment domain	Question 10, range 0 to 3.	N/A – partial assessments cannot be saved.
		DLQI total score	DLQI total score: sum of all six DLQI domain scores, range 0 to 30.	N/A – partial assessments cannot be saved.
		Change from baseline in DLQI	Change from baseline: observed DLQI score – baseline DLQI score	Missing if baseline or observed value is missing.
		4-point improvement in DLQI total score in subgroup of patients with baseline ≥4	Change from baseline ≤4 and baseline ≥4	Missing if baseline is missing or <4 or observed value is missing.
		DLQI (0,1)	A DLQI (0,1) response is defined as a post-baseline DLQI total score of 0 or 1	Missing if the DLQI total score is missing

Measure	Description	Variable	Derivation / Comment	Imputation Approach if Missing Components
Work Productivity and Activity Impairment: Atopic Dermatitis (WPAI-AD)	The Work Productivity and Activity Impairment Questionnaire–Atopic Dermatitis (WPAI-AD) records impairment due to AD during the past 7 days. The WPAI-AD consists of 6 items grouped into 4 domains: absenteeism (work time missed), presenteeism (impairment at work/reduced on-the-job effectiveness), work productivity loss (overall work impairment/absenteeism plus presenteeism), and activity impairment. Scores are calculated as impairment percentages (Reilly et al. 1993), with higher scores indicating greater impairment and less productivity.	Employment status	Question (Q)1	Single item, missing if missing.
		Change in employment status	Employed at baseline and remained employed: Q1 = 1 at post-baseline visit and at baseline visit. Not employed at baseline and remain unemployed: Q1 = 0 at post-baseline visit and at baseline visit.	Missing if baseline or observed value is missing.
		Percentage of absenteeism	Percent work time missed due to problem: (Q2/(Q2 + Q4))*100	If Q2 or Q4 is missing, then missing.
		Change from baseline in absenteeism	Change from baseline: observed absenteeism – baseline absenteeism	Missing if baseline or observed value is missing.
		Percentage of presenteeism	Percent impairment (reduced productivity while at work) while working due to problem: (Q5/10)*100	If Q5 is missing, then missing.
		Change from baseline in presenteeism	Change from baseline: observed presenteeism – baseline absenteeism	Missing if baseline or observed value is missing.
		Overall work impairment	Percent overall work impairment (combines absenteeism and presenteeism) due to problem: (Q2/(Q2+Q4) + [(1-Q2/(Q2+Q4))*(Q5/10)])*100	If Q2, Q4, or Q5 is missing, then missing.

Measure	Description	Variable	Derivation / Comment	Imputation Approach if Missing Components
		Change from baseline in work impairment	Change from baseline: observed work impairment – baseline work impairment	Missing if baseline or observed value is missing.
		Percentage of impairment in activities	Percent activity impairment (performed outside of work) due to problem: (Q6/10)*100	If Q6 is missing, then missing.
		Change from baseline in impairment in activities	Change from baseline: observed impairment in activities – baseline impairment in activities	Missing if baseline or observed value is missing.
European Quality of Life–5 Dimensions–5 Levels (EQ-5D-5L)	The European Quality of Life–5 Dimensions–5 Levels (EQ-5D-5L) is a standardized measure of health status that provides a simple, generic measure of health for clinical and economic appraisal. The EQ-5D-5L consists of 2 components: a descriptive system of the respondent's health and a rating of his or her current health state using a 0 to 100 mm VAS. The descriptive system comprises the following 5 dimensions: mobility, self-care, usual activities, pain/discomfort, and anxiety/depression. Each dimension has 5 levels: no problems, slight problems, moderate problems, severe problems, and extreme problems. The respondent is asked to indicate his or her health state by ticking (or placing a cross) in the box associated with the most appropriate statement in each of the 5 dimensions. It should be noted that the numerals 1 to 5 have no arithmetic properties and should not be used as an ordinal score. The VAS records the respondent's self-rated health on a vertical VAS where the endpoints are	▪ EQ-5D mobility ▪ EQ-5D self-care ▪ EQ-5D usual activities ▪ EQ-5D pain/ discomfort ▪ EQ-5D anxiety/ depression	Five health profile dimensions, each dimension has 5 levels: 1 = no problems 2 = slight problems 3 = moderate problems 4 = severe problems 5 = extreme problems It should be noted that the numerals 1 to 5 have no arithmetic properties and should not be used as a primary score.	Each dimension is a single item, missing if missing.
		EQ-5D VAS	Single item. Range 0 to 100. 0 represents "worst health you can imagine" 100 represents "best health you can imagine"	Single item, missing if missing.
		Change from baseline in EQ-5D VAS	Change from baseline: observed EQ-5D VAS score – baseline EQ-5D VAS score	Missing if baseline or observed value is missing.
		EQ-5D-5L UK Population-based index score (health	Derive EQ-5D-5L UK Population-based index score according to the link by using the	N/A – partial assessments cannot be saved

Measure	Description	Variable	Derivation / Comment	Imputation Approach if Missing Components
	labeled "best imaginable health state" and "worst imaginable health state." This information can be used as a quantitative measure of health outcome. The EQ-5D-5L health states, defined by the EQ-5D-5L descriptive system, may be converted into a single summary index by applying a formula that essentially attaches values (also called weights) to each of the levels in each dimension (Herdman et al. 2011; EuroQol Group 2015 [WWW]).	state index)	UK algorithm to produce a patient-level index score between -0.59 and 1.0 (continuous variable).	on the eCOA tablet.
		Change from baseline in EQ-5D-5L UK Population-based index score	Change from baseline: observed EQ-5D-5L UK score – baseline EQ-5D-5L UK score	Missing if baseline or observed value is missing.
		EQ-5D-5L US Population-based index score (health state index)	Derive EQ-5D-5L US Population-based index score according to the link by using the US algorithm to produce a patient-level index score between -0.11 and 1.0 (continuous variable).	N/A – partial assessments cannot be saved on the eCOA tablet.
		Change from baseline in EQ-5D-5L US Population-based index score	Change from baseline: observed EQ-5D-5L US score – baseline EQ-5D-5L US score	Missing if baseline or observed value is missing.
PROMIS Itch Questionnaire (PIQ)	PIQ – Itch Interference: consists of 8 items assessing the impact of itch on various aspects of life. PIQ – Activity and Clothing: consists of 8 items assessing activity and clothing related quality of life impairment from itch in adults "in the past 7 days". PIQ – Mood and Sleep: consists of 8 items assessing mood and sleep related quality of life impairment from itch and impact of itch "in the past 7 days". PIQ – Scratching Behavior: consists of 5 items assessing quality of life impairment from scratching behavior and the physical	PIQ – Itch Interference[a]	8 items. Each range 1 to 5. 1=Never, 2=Rarely, 3=Sometimes, 4=Often, 5=Almost always (continuous variable). Total raw scores are converted to T-Scores with higher scores representing greater impact because of itch.	Calculation is made by HealthMeasures Scoring Service, powered by Assessment Center[SM]
		Change from baseline in PIQ – Itch Interference	Change from baseline: observed PIQ Itch Interference score – baseline PIQ Itch Interference score	Missing if Baseline or observed value is missing.
		PIQ – Activity and Clothing	8 items. Each range 1 to 5. 1=Never, 2=Rarely,	Score is calculated by

Measure	Description	Variable	Derivation / Comment	Imputation Approach if Missing Components
	manifestations of itch in adults "in the past 7 days".		3=Sometimes, 4=Often, 5=Almost always (continuous variable). Total raw scores are converted to T-Scores with higher scores representing greater impact because of itch.	HealthMeasures Scoring Service, powered by Assessment Center[SM]
		Change from baseline in PIQ – Activity and Clothing	Change from baseline: observed PIQ Activity and Clothing score – baseline PIQ Activity and Clothing score	Missing if Baseline or observed value is missing.
		PIQ – Mood and Sleep	8 items. Each range 1 to 5. 1=Never, 2=Rarely, 3=Sometimes, 4=Often, 5=Almost always (continuous variable). Total raw scores are converted to T-Scores with higher scores representing greater impact because of itch.	Score is calculated by HealthMeasures Scoring Service, powered by Assessment Center[SM]
		Change from baseline in PIQ – Mood and Sleep	Change from baseline: observed PIQ Mood and Sleep score – baseline PIQ Mood and Sleep score	Missing if Baseline or observed value is missing.
		PIQ – Scratching Behavior	5 items. Each range 1 to 5. Response options for the frequency of scratching behaviors: 1=Never, 2=Rarely, 3=Sometimes, 4=Often, 5=Almost always The response options for the worry related to scratching items: 1=Not at all, 2=A little bit, 3=Somewhat, 4=Quite a bit, 5=Very much (continuous variable). Total raw scores are converted to T-Scores with higher	Score is calculated by HealthMeasures Scoring Service, powered by Assessment Center[SM]

Measure	Description	Variable	Derivation / Comment	Imputation Approach if Missing Components
			scores representing more scratching behavior.	
		Change from baseline in PIQ – Scratching Behavior	Change from baseline: observed PIQ Scratching Behavior score – baseline PIQ Scratching Behavior score	Missing if Baseline or observed value is missing.
PROMIS – Sleep Related Impairment	The Sleep Related Impairment Short Form within the PROMIS bank consists of 8 items measuring self-reported perceptions of alertness, sleepiness, and tiredness during usual waking hours, and the perceived functional impairments during wakefulness associated with sleep problems or impaired alertness "in the past 7 days". Response options range from 1=Not at all; 2=A little bit; 3=Somewhat; 4=Quite a bit; to 5=Very much.	PROMIS sleep related impairment	8 items. Each range 1 to 5. 1=Not at all; 2=A little bit; 3=Somewhat; 4=Quite a bit; to 5=Very much (continuous variable). Total raw scores are converted to T-Scores with higher scores representing greater sleep impairment.	Score is calculated by HealthMeasures Scoring Service, powered by Assessment Center[SM]
		Change from baseline in PROMIS sleep related impairment	Change from baseline: observed PROMIS Sleep Related Impairment score – baseline PROMIS Sleep Related Impairment score	Missing if Baseline or observed value is missing.
Neuro-QoL – Cognitive Function	The Cognitive Function Short Form domain within Neuro-QoL bank consists of 8-items measuring Executive Function (perceived difficulties in applications of mental health function related to planning, organizing, calculating, remembering and learning) "in the past 7 days" and General Concerns (perceived difficulties in everyday cognitive abilities such as memory, attention, and decision making) using the lead-in phrase "how much difficulty do you currently have…". The response options for the Executive Function items range from 1=Very Often (several times a day); 2=Often (once a day); 3=Sometimes (2-3 times); 4=Rarely (once); to	Neuro-QoL – Cognitive Function	The total raw scores are converted to T-Scores with higher scores indicating better (desirable) self-reported health.	Score is calculated by HealthMeasures Scoring Service, powered by Assessment Center[SM]
		Change from baseline in Neuro-QoL – Cognitive Function	Change from baseline: observed Neuro-QoL – Cognitive Function score – baseline Neuro-QoL – Cognitive Function score	Missing if Baseline or observed value is missing.

Measure	Description	Variable	Derivation / Comment	Imputation Approach if Missing Components
	5=Never). The response options for the General Concerns items range from 1=Cannot do; 2=A lot; 3=Somewhat; 4=A little, to 5=None.			
Patient Benefit Index (PBI)	The Patient Benefit Index (PBI) measures patient-defined treatment objectives and benefits. It consists of 2 questionnaires. Before therapy, patients complete the standardized "Patient Needs Questionnaire" (PNQ) indicating individual importance of treatment objectives. This reflects their personal preferences with respect to therapeutic benefit. During the study patients rate the extent to which the treatment objectives have been achieved in the "Patient Benefit Questionnaire" (PBQ). Subscales of the Patient Benefit Index are: • Reducing social impairments subscale score: item 11, 13, 14, 15, 16, 17 • Reducing psychological impairments subscale score: item 6, 7, 9, 10, 12 • Reducing impairments due to therapy subscale score: item 18, 19, 20, 21 • Reducing physical impairments subscale score: item 1, 2, 3, 4, 5 • Having confidence in healing subscale score: item 8, 22, 23	• Patient Benefit Index (PBI) global score • Subscale scores	25 items. 0=not at all, 1=somewhat; 2=moderately; 3=quite; 4=very; PBI global score is calculated for each patient by weighing the achievement values of the treatment objectives by their importance to the individual patient. $$PBI = \sum_{i=1}^{k} \frac{PNQ_i}{\sum_{i=1}^{k} PNQ_i} PBQ_i$$ For score calculation, both "does/did not apply" and question unanswered will be treated as missing values. The global score is calculated using only these item pairs, for which the patient has given a response other than "does/did not apply to me" in both PNQ and PBQ. Subscale scores are calculated in the same manner as the global score.	PBI global score and subscales may only be computed if the patient has provided at least 75% valid data in each of the PNQ and PBQ respectively. In this context, the responses "not at all" and "does/did not apply" count as valid data. Thus a treatment goal is regarded missing if the patient has not responded to the item in the PNQ and/or in the PBI.

Measure	Description	Variable	Derivation / Comment	Imputation Approach if Missing Components
		• PBI global score ≥1	PBI global score at least 1	Missing if observed value is missing

[a] PIQ – Itch Interference here is the same as PIQ – General in the protocol.

Table JAIY.6.7. Description of Health Outcomes and Quality-of-Life Measures Analyses

Measure	Variable	Analysis Method (Section 6.2.3)	Population (Section 6.2.1)	Comparison/Time Point	Analysis Type
Itch Numeric Rating Scale (NRS)	• Itch NRS score • Change from baseline in Itch NRS score	MMRM	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 1, …, 16	Secondary Analysis
		ANCOVA using mLOCF	ITT	Bari 4 mg or Bari 2 mg vs PBO; Day 2, Week 1,…, 16	Secondary Analysis at Day 2; Sensitivity Analysis for others
		pMI	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 1, …, 16	Sensitivity Analysis
	Percent change from baseline Itch score	MMRM	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 1, …,16	Secondary Analysis
		ANCOVA using mLOCF	ITT	Bari 4 mg or Bari 2 mg vs PBO; Day 2, Week 1, …, 16	Secondary Analysis at Day 2; Sensitivity Analysis for others
	Proportion of patients achieving a 4-point improvement in Itch NRS in subgroup of patients who had baseline Itch NRS ≥4	Logistic regression using NRI	ITT	Bari 4 mg or Bari 2 mg vs PBO; Day 2, Week 1,…, 16	Key Secondary Analysis at Day 2, Week 1, 2 and 16
			PPS	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Sensitivity analysis
		Logistic regression using pMI and Tipping Point	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Sensitivity analysis
	Number of Itch-free (Itch NRS = 0) Days	Two-sample t-test	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 0 to Week 4, …, Week 12 to 16	Exploratory Analysis
Skin Pain Numeric Rating Scale (NRS)	• Skin Pain NRS score • Change from baseline in Skin Pain NRS score	MMRM	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 1,…, 16	Key Secondary Analysis at Week 16
			PPS	Bari 4 mg or Bari 2 mg vs PBO; Week 1,…, 16	Sensitivity analysis

Measure	Variable	Analysis Method (Section 6.2.3)	Population (Section 6.2.1)	Comparison/Time Point	Analysis Type
		ANCOVA using mLOCF	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 1,…, 16	Sensitivity Analysis
		pMI	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 1,…, 16	Sensitivity analysis
	Number of Skin Pain-free (Skin pain NRS = 0) Days	Two-sample t-test	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 0 to Week 4,…, Week 12 to 16	Exploratory Analysis
	Proportion of patients achieving a 4-point improvement in Skin Pain NRS in subgroup of patients with baseline Skin Pain NRS ≥4	Logistic regression using NRI	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 1,…, 16	Other Secondary Analysis
Atopic Dermatitis Sleep Scale (ADSS)	• ADSS item scores • Change from baseline in ADSS item scores	MMRM	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 1,…, 16	Key Secondary Analysis at Week 1 and 16
			PPS	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Sensitivity analysis
		ANCOVA using mLOCF	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 1,…, 16	Sensitivity Analysis
		pMI	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Sensitivity analysis
Patient-Oriented Eczema Measure (POEM)	• POEM score • Change from baseline in POEM score	MMRM	ITT	Bari 4 mg or Bari 2 mg vs PBO; at each post-baseline visit	Secondary Analysis
		ANCOVA using mLOCF	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Sensitivity Analysis
	Proportion of patients achieving a 4-point improvement in subgroup of patients with baseline POEM ≥ 4	Logistic regression using NRI	ITT	Bari 4 mg or Bari 2 mg vs PBO; at each post-baseline visit	Secondary Analysis
Patient Global Impression of Severity–Atopic Dermatitis (PGI-S-AD)	• PGI-S-AD score • Change from baseline in PGI-S-AD score	MMRM	ITT	Bari 4 mg or Bari 2 mg vs PBO; at each post-baseline visit	Secondary Analysis
		ANCOVA using mLOCF	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Sensitivity Analysis

Measure	Variable	Analysis Method (Section 6.2.3)	Population (Section 6.2.1)	Comparison/Time Point	Analysis Type
Hospital Anxiety Depression Scale (HADS)	• HADS domain scores • Change from baseline in HADS domain	MMRM	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 2, 4, 8,16	Secondary Analysis
		ANCOVA using mLOCF	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Sensitivity Analysis
Dermatology Life Quality Index (DLQI)	• DLQI total score • Change from baseline in DLQI • Observed and change from baseline in domain scores -Symptoms and feelings -Daily activities -Leisure -Work and school -Personal relationships -Treatment	MMRM	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Secondary Analysis
		ANCOVA using mLOCF	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Sensitivity Analysis
	Proportion of patients achieving a 4-point improvement in DLQI total score in subgroup of patients with baseline DLQI ≥4	Logistic regression using NRI	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 1, 2, 4, 8, 16	Other Secondary Analysis
	Proportion of patients achieving DLQI (0,1)	Logistic regression using NRI	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 1, 2, 4, 8, 16	Other Secondary Analysis
Work Productivity and Activity Impairment: Atopic Dermatitis (WPAI-AD)	Observed and Change from baseline in employment status	Descriptive statistics (observed)	ITT	No comparisons; Week 16	Secondary Analysis
	Observed and Change from baseline in: • absenteeism • presenteeism • overall work impairment • impairment in activities	MMRM	ITT	Bari 4 mg or Bari 2 mg vs PBO; at each post-baseline visit	Secondary Analysis
		ANCOVA using mLOCF	ITT	Bari 4 mg or Bari 2 mg vs PBO; at each post-baseline visit	Sensitivity Analysis

Measure	Variable	Analysis Method (Section 6.2.3)	Population (Section 6.2.1)	Comparison/Time Point	Analysis Type
European Quality of Life–5 Dimensions–5 Levels (EQ-5D-5L)	Observed values in • EQ-5D mobility • EQ-5D self-care • EQ-5D usual activities • EQ-5D pain/ discomfort • EQ-5D anxiety/ depression	Logistic Regression using NRI	ITT	Bari 4 mg or Bari 2 mg vs PBO: at each post-baseline visit	Exploratory Analysis
	Observed and Change from baseline in • EQ-5D VAS • EQ-5D-5L UK Population-based index score • EQ-5D-5L US Population-based index score	MMRM	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Secondary Analysis
		ANCOVA using mLOCF	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Sensitivity Analysis
PROMIS	Observed and Change from baseline in: • PIQ – Itch Interference score • PIQ – Activity and Clothing score • PIQ – Mood and Sleep score • PIQ – Scratching Behavior score • PROMIS – Sleep-Related Impairment score	MMRM	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 1, 2, 4, 8, 12, 16	Exploratory Analysis
		ANCOVA using mLOCF	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 1, 2, 4, 8, 12, 16	Sensitivity Analysis
Neuro-QoL – Cognitive Function score	Observed and Change from baseline in: Neuro-QoL – Cognitive Function score	ANCOVA using mLOCF	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Exploratory Analysis

Measure	Variable	Analysis Method (Section 6.2.3)	Population (Section 6.2.1)	Comparison/Time Point	Analysis Type
PBI	• PBI global score • Reducing social impairments subscale score • Reducing psychological impairments subscale score • Reducing impairments due to therapy subscale score • Reducing physical impairments subscale score • Having confidence in healing subscale score	ANOVA[a]	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Exploratory Analysis
PBI	Proportion of patients achieving PBI global score ≥1 at Week 16	Logistic regression using NRI	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Exploratory Analysis

Abbreviations: ANCOVA = analysis of covariance; ANOVA= analysis of variance; Bari = baricitinib; ITT = intent-to-treat; mLOCF = modified last observation carried forward; MMRM = mixed model repeated measures; NRI = nonresponder imputation; PBO = placebo; pMI=placebo multiple imputation; PPS = per protocol set.

Notes: for all other post-baseline visits not mentioned in the table, but collected for the measures as specified in the protocol, the analyses are made as exploratory analyses.

Notes: all the key secondary are performed for Japan population. Other secondary and exploratory analyses may be performed for Japan population.

[a] ANOVA model includes region, baseline disease severity (IGA) and treatment as factors in the model.

6.14. Bioanalytical and Pharmacokinetic/Pharmacodynamic Methods

Pharmacokinetic (PK), Pharmacodynamic (PD) and Biomarker analyses to address secondary and exploratory objectives of this study will be described by Lilly in separate PK/PD and Biomarker analysis plans.

6.15. Safety Analyses

The general methods used to summarize safety data, including the definition of baseline and postbaseline are described in Section 6.2.

Safety analyses will include data from first dose of the study treatment to after rescue, unless otherwise stated, and patients will be analyzed according to the investigational product to which they were randomized at Visit 2. A sensitivity approach to the safety analyses will use data censored at last dose of the study drug for patients rescued to systemic therapy. These analyses will be conducted for treatment-emergent adverse events (TEAEs), serious adverse events (SAEs), AEs leading to permanent study drug discontinuation and special topics excluding deaths and malignancies. Additional analyses may be conducted using data after rescue to systemic therapy for some safety topics such as systemic TEAEs, and SAEs. Safety analyses will use the safety population defined in Section 6.2.1.

Safety topics that will be addressed include the following: AEs including TEAEs and SAEs, clinical laboratory evaluations, vital signs and physical characteristics, Columbia Suicide Severity Rating Scale (C-SSRS), the Self-Harm Supplement Form, safety in special groups and circumstances, including adverse events of special interest (AESI) (see Section 6.15.5), and investigational product interruptions.

Unless otherwise specified, by-visit summaries will include planned on-treatment visits. For tables that summarize events (such as AEs, categorical lab abnormalities, shift to maximum value), post-last dose follow-up data will be included. Follow-up data is defined as all data occurring up to 30 days (planned maximum follow-up time) after last dose of treatment including rescue, regardless of study period.

For selected safety assessments other than events, descriptive statistics may be presented for the last measure observed during post-treatment follow-up (up to 30 days after the last dose of treatment including rescue, regardless of study period).

6.15.1. Extent of Exposure

Duration of exposure (in days) will be calculated as follows:

- Duration of exposure to investigational product (including exposure after the initiation of rescue therapy): *date of last dose of study drug including rescue – date of first dose of study drug + 1*.

Last dose of study drug including rescue is calculated as last date on study drug. See the compound level safety standards for more details.

Total patient-years (PY) of exposure to study drug will be reported for each treatment group for overall duration of exposure. Descriptive statistics will be provided for patient-days of exposure and the frequency of patients falling into different exposure ranges in addition to cumulative exposures will be summarized.

Exposure ranges will be summarized as follows:

- ≥28 days, ≥56 days, ≥84 days, and ≥112 days
- >0 to <28 days, ≥28 days to <56 days, ≥56 days to <84 days, ≥84 days to <112 days, and ≥112 days

Overall exposure for a treatment group will be summarized in total PY which is calculated according to the following formula:

- *Exposure in PY (PYE) = sum of duration of exposure in days* (for all patients in treatment group) */ 365.25*

6.15.2. Adverse Events

Adverse events are recorded in the eCRFs. Each AE will be coded to system organ class (SOC) and preferred term (PT) using the Medical Dictionary for Regulatory Activities (MedDRA) version that is current at the time of database lock. Severity of AEs is recorded as mild, moderate, or severe.

A TEAE is defined as an event that either first occurred or worsened in severity after the first dose of study treatment and on or prior to the last visit date during the analysis period. The analysis period is defined as the treatment period plus up to 30 days off-drug follow-up time.

Adverse events are classified based upon the MedDRA PT. The MedDRA Lowest Level Term (LLT) will be used in defining which events are treatment-emergent. The maximum severity for each LLT during the baseline period up to first dose of the study medication will be used as baseline. If an event with missing severity is preexisting during the baseline period, and persists during the treatment period, then the baseline severity will be considered mild for determining treatment-emergence (that is, the event is treatment-emergent if the severity is coded moderate or severe postbaseline and not treatment-emergent if the severity is coded mild postbaseline). If an event occurring postbaseline has a missing severity rating, then the event is considered treatment-emergent unless the baseline rating is severe, in which case the event is not treatment-emergent. The day and time for events where onset is on the day of the first dose of study treatment will both be used to distinguish between pretreatment and posttreatment to derive treatment-emergence. Should there be insufficient data for AE start date to make this comparison (for example, the AE start year is the same as the treatment start year, but the AE start month and day are missing), the AE will be considered treatment-emergent.

In general, summaries will include the number of patients in the safety population (N), frequency of patients reporting the event (n), and relative frequency (that is, percentage; n/N*100). For any events that are gender-specific based on the displayed PT, the denominator used to compute the percentage will only include patients from the appropriate gender.

In an overview table, the number and percentage of patients in the safety population who experienced death, an SAE, any TEAE, discontinuation from the study due to an AE, permanent discontinuation from study drug due to an AE, or a severe TEAE will be summarized by treatment group.

The number and percentage of patients with TEAEs will be summarized by treatment group in 2 formats by MedDRA PT:

- nested within SOC with decreasing frequency in SOC, and events ordered within each SOC by decreasing frequency in the baricitinib 4-mg group;
- with events ordered by decreasing frequency in the baricitinib 4-mg group.

6.15.2.1. Common Adverse Events

Common TEAEs are defined as TEAEs that occurred in ≥2% (before rounding) of patients in any treatment group including placebo. The number and percentage of patients with common TEAEs will be summarized by treatment using MedDRA PT ordered by decreasing frequency in the baricitinib 4-mg group.

The number and percentage of patients with TEAEs will be summarized by maximum severity by treatment using MedDRA PT ordered by decreasing frequency in the baricitinib 4-mg group for the common TEAEs. For each patient and TEAE, the maximum severity for the MedDRA level being displayed is the maximum postbaseline severity observed from all associated LLTs mapping to that MedDRA PT.

6.15.2.2. Serious Adverse Event Analyses

Consistent with the International Conference on Harmonisation (ICH) E2A guideline (1994) and 21 Code of Federal Regulations (CFR) 312.32 (a) (2010), a SAE is any AE that results in any one of the following outcomes:

- death
- initial or prolonged inpatient hospitalization
- a life-threating experience (that is, immediate risk of dying)
- persistent or significant disability/incapacity
- congenital anomaly/birth defect

The number and percentage of patients who experienced any SAE will be summarized by treatment using MedDRA PT nested within SOC. Events will be ordered by decreasing frequency in the baricitinib 4-mg group within decreasing frequency in SOC. The SAEs will also be summarized by treatment using MedDRA PT without SOC.

An individual listing of all SAEs will be provided. A listing of deaths, if any, regardless of when they occurred during the study, will also be provided.

6.15.2.3. Other Significant Adverse Events

Other significant AEs to be summarized will provide the number and percentage of patients who:

- permanently discontinued study drug because of an AE or death

- temporarily interrupted study drug because of AE

by treatment using MedDRA PT nested within SOC. Events will be ordered by decreasing frequency in the baricitinib 4-mg group within decreasing frequency in SOC.

A summary of temporary interruptions of study drug will also be provided, showing the number of patients who experienced at least one temporary interruption and the number of temporary interruptions per patient with an interruption. Further, the duration of each temporary interruption (in days), the cumulative duration of dose interruption (in days) using basic descriptive statistics and the reason for dose interruption will be provided.

A listing of all AEs leading to permanent discontinuation from the study drug or from the study will be provided. A listing of all temporary study drug interruptions, including interruptions for reasons other than AEs, will be provided.

6.15.2.4. Criteria for Notable Patients

Patient narratives will be provided for all patients who experience certain "notable" events prior to data cutoff for the submission. See compound level safety standards for list of criteria.

6.15.3. Clinical Laboratory Evaluation

For the categorical laboratory analyses (shift and treatment emergent), the analysis period is defined as the treatment period plus up to 30 days off-drug follow-up time. The analysis period for the continuous laboratory analyses (e.g., change from baseline by time point) is defined as the treatment period excluding off-drug follow-up time.

All laboratory tests will be presented using the International Système (SI) and US conventional (CN) units. The performing central laboratory reference ranges will be used to define the low and high limits. Results pertaining to the 4 key hepatic laboratory assessments (alanine aminotransferase [ALT], aspartate aminotransferase [AST], total bilirubin, and alkaline phosphatase [ALP]) will be included as a separate analysis to address the risk of liver injury as a special safety topic (see Section 6.15.5.1).

There is one special circumstance for laboratory values to be derived based on regularly scheduled, protocol-specified analytes. The low-density lipoprotein/high-density lipoprotein (LDL/HDL) ratio will be derived as the ratio of LDL cholesterol to HDL cholesterol. There are no central lab reference ranges for the LDL/HDL ratio.

The following will be conducted for the laboratory analytes collected quantitatively:

- Box plots: Values at each visit (starting from randomization) and change from last baseline to each visit and to last postbaseline measure will be displayed in box plots for patients who have both a baseline and at least 1 postbaseline visit. The last non-missing observation in the treatment period will be used as the last observation. Individual measurements outside of reference limits will also be displayed using distinct symbols overlaying the box plot. Original-scale data will be used for the display but for some analytes (for example, immunoglobulins) a logarithmic scale may be used to aid in viewing the measures of central tendency and dispersion. Unplanned measurements will

be excluded. Descriptive summary statistics will be included below the box plot along with p-values resulting from between treatment comparison in change from last baseline to last observation. An ANCOVA model with explanatory term for treatment and the baseline value as a covariate will be used. These box plots will be used to evaluate trends over time and to assess a potential impact of outliers on central tendency summaries.

- Treatment-emergent high/low analyses: The number and percentage of patients with treatment-emergent high and low laboratory results at any time will be summarized by treatment group. Planned and unplanned measurements will be included. A treatment-emergent **high** result is defined as a change from a value less than or equal to the high limit at all baseline visits to a value greater than the high limit at any time during the treatment period. A treatment-emergent **low** result is defined as a change from a value greater than or equal to the low limit at all baseline visits to a value less than the low limit at any time during the treatment period. The Fisher's exact test will be used for the treatment comparisons. Number at risk (NAR) for the treatment-emergent high result is defined as the number of patients with a value less than or equal to the high limit at all baseline visits. NAR for the treatment-emergent low result is defined as the number of patients with a value greater than or equal to the low limit at all baseline visits for the treatment-emergent low result.

A listing of abnormal findings will be provided for laboratory analyte measurements, including qualitative measures. The listing will include but not limited to patient ID, treatment group, laboratory collection date, analyte name, and analyte finding. If needed by the safety physician/scientist, for analytes measured qualitatively, the number and percentage of patients with treatment-emergent abnormal laboratory results at any time will be summarized by treatment. Planned and unplanned measurements will be included. A treatment-emergent abnormal result is defined as a change from normal at all baseline visits to abnormal at any time postbaseline.

Note that additional analyses of certain laboratory analytes will be discussed within sub-sections of Section 6.15.5 pertaining to Special Safety topics (Section 6.15.5.1 for hepatic analytes, Section 6.15.5.2 for analytes related to hematological changes, Section 6.15.5.3 for analytes related to lipids, Section 6.15.5.4 for analytes related to renal function, and Section 6.15.5.5 for CPK).

6.15.4. Vital Signs and Other Physical Findings

For the treatment-emergent categorical analyses (shift and treatment emergent), the analysis period is defined as the treatment period plus up to 30 days off-drug follow-up time. The analysis period for the continuous analyses (e.g., change from baseline by time point) is defined as the treatment period excluding off-drug follow-up time.

Vital signs and physical characteristics include systolic blood pressure (SBP), diastolic blood pressure (DBP), pulse, weight, and BMI. Original-scale data will be analyzed. When these parameters are analyzed as continuous numerical variables, unplanned measurements will be

excluded. When these parameters are analyzed as categorical outcomes and/or treatment-emergent abnormalities, planned and unplanned measurements will be included.

The planned analyses described for the laboratory analytes in Section 6.15.3 will be used to analyze the vital signs and physical characteristics.

Table JAIY.6.8 defines the low and high baseline values as well as the criteria used to define treatment-emergence based on post-baseline values. The blood pressure and pulse rate criteria are consistent with the document *Selected Reference Limits for Pulse/Heart Rate, Arterial Blood Pressure (Including Orthostasis), and Electrocardiogram Numerical Parameters for Use in Analyses of Phase 2-4 Clinical Trials Version 1.3* approved on April 29, 2015 as recommended by the Lilly Cardiovascular Safety Advisory Committee (CVSAC).

Table JAIY.6.8. Categorical Criteria for Abnormal Treatment-Emergent Blood Pressure and Pulse Measurement, and Categorical Criteria for Weight Changes for Adults

Parameter (Units of Measure)	Low	High
Systolic Blood Pressure (mm Hg)	≤90 (low limit) and decrease from lowest value during baseline ≥20 if >90 at each baseline visit	≥140 (high limit) and increase from highest value during baseline ≥20 if <140 at each baseline visit
Diastolic Blood Pressure (mm Hg)	≤50 (low limit) and decrease from lowest value during baseline ≥10 if >50 at each baseline visit	≥90 (high limit) and increase from highest value during baseline ≥10 if <90 at each baseline visit
Pulse (beats per minute)	<50 (low limit) and decrease from lowest value during baseline ≥15 if ≥50 at each baseline visit	>100 (high limit) and increase from highest value during baseline ≥15 if ≤100 at each baseline visit
Weight (kilograms)	(Loss) decrease ≥7% from lowest value during baseline	(Gain) increase ≥7% from highest value during baseline

6.15.5. Special Safety Topics, including Adverse Events of Special Interest

In addition to general safety parameters, safety information on specific topics of special interest will also be presented. Additional special safety topics may be added as warranted. The topics outlined in this section include the protocol-specified AESI.

In general, for topics regarding safety in special groups and circumstances, patient profiles and/or patient listings, where applicable, will be provided when needed to allow medical review of the time course of cases/events, related parameters, patient demographics, study drug treatment and meaningful concomitant medication use. In addition to the safety topics for which provision or review of patient data is specified, these will be provided when summary data are insufficient to permit adequate understanding of the safety topic.

6.15.5.1. Abnormal Hepatic Tests

Analyses for abnormal hepatic tests will involve 4 laboratory analytes: ALT, AST, total bilirubin, and ALP. In addition to the analyses described in Section 6.15.3, this section describes specific analyses for this topic.

First, the number and percentage of patients with the following abnormal elevations in hepatic laboratory tests at any time will be summarized between treatment groups:

- The percentages of patients with an ALT measurement ≥3×, 5×, and 10× the central laboratory upper limit of normal (ULN) during the treatment period will be summarized for all patients with a postbaseline value and for subsets based on various levels of baseline.
    - The analysis of 3× ULN will contain 4 subsets: patients whose non-missing maximum baseline value is ≤1× ULN, patients whose maximum baseline is >1× ULN but <3× ULN, patients whose maximum baseline value is ≥3× ULN, and patients whose baseline values are missing.
    - The analysis of 5× ULN will contain 5 subsets: patients whose non-missing maximum baseline value is ≤1× ULN, patients whose maximum baseline is >1× ULN but <3× ULN, patients whose maximum baseline is ≥3× ULN but <5× ULN, patients whose maximum baseline value is ≥5× ULN, and patients whose baseline values are missing.
    - The analysis of 10× ULN will contain 6 subsets: patients whose non-missing maximum baseline value is ≤1× ULN, patients whose maximum baseline is >1× ULN but <3× ULN, patients whose maximum baseline is ≥3× ULN but <5× ULN, patients whose maximum baseline is ≥5× ULN but <10× ULN, patients whose maximum baseline value is ≥10× ULN, and patients whose baseline values are missing.
- The percentages of patients with an AST measurement ≥3×, 5×, and 10× the central laboratory ULN during the treatment period will be summarized for all patients with a postbaseline value and for subsets based on various levels of baseline. Analyses will be constructed as described above for ALT.
- The percentages of patients with a total bilirubin measurement ≥2× the central laboratory ULN during the treatment period will be summarized for all patients with a postbaseline value and subset into 4 subsets: patients whose non-missing maximum baseline value is ≤1× ULN, patients whose maximum baseline is >1× ULN but <2× ULN, patients whose maximum baseline value is ≥2× ULN, and patients whose baseline values are missing.
- The percentages of patients with an ALP measurement ≥1.5× the central laboratory ULN during the treatment period will be summarized for all patients with a postbaseline value and subset into 4 subsets: patients whose non-missing maximum baseline value is ≤1× ULN, patients whose maximum baseline is >1× ULN but <1.5× ULN, patients whose maximum baseline value is ≥1.5× ULN, and patients whose baseline values are missing.

Information collected from additional hepatic safety data collection forms will be provided in patient profiles.

Second, to further evaluate potential hepatotoxicity, an Evaluation of Drug-Induced Serious Hepatotoxicity (eDISH) plot using maximum postbaseline ALT divided by ULN vs. maximum postbaseline total bilirubin divided by ULN will be created that includes all patients from the safety populations for the studies included in the submission (any phase, any medication). Each subject with at least 1 postbaseline ALT and total bilirubin contributes 1 point to the plot. The measurements do not need to be taken at the same blood draw. Symbols may be used to indicate randomized treatment.

When criteria are met for hepatic evaluation and completion of the hepatic safety CRF, investigators are required to answer a list of questions (see Compound level safety standards). A listing of the collected information will be generated together with a graphical patient profile. This includes demographics, disposition, and a display of study drug exposure, AEs, medications, and the liver-related measurements over time will be provided for these patients and any additional patients meeting ALT or AST measurement greater than or equal to 5× ULN (on a single measurement) or ALP measurement greater than or equal to 2× ULN (on a single measurement).

6.15.5.2. Hematologic Changes

Hematologic changes will be defined based on clinical laboratory assessments. Common Terminology Criteria for Adverse Events (CTCAEs) will be applied for selected laboratory tests and are described in the compound level safety standards. These CTCAE grading schemes are consistent with both Version 3.0 and Version 4.03 of the CTCAE guidelines (CTCAE 2003, 2010).

Treatment-emergent laboratory abnormalities occurring at any time during the treatment period and shift tables of baseline to maximum grade during the treatment period will be tabulated. Planned and unplanned measurements will be included. Treatment-emergence will be characterized using the following 5 criteria (as appropriate to the grading scheme):

- any increase in postbaseline CTCAE grade from worst baseline grade
- increase to Grade 1 or above at worst postbaseline
- increase to Grade 2 or above at worst postbaseline
- increase to Grade 3 or above at worst postbaseline
- increase to Grade 4 at worst postbaseline.

Shift tables will show the number and percentage of patients based on baseline to maximum during the treatment period, with baseline depicted by the most extreme grade during the baseline period. With each shift table, a shift table summary displaying the number and percentage of patients with maximum postbaseline results will be presented by treatment group for each treatment period within the following categories:

- decreased: postbaseline category < baseline category
- increased: postbaseline category > baseline category

- same: postbaseline category = baseline category

A laboratory-based treatment-emergent outcome related to increased platelet count will be summarized in similar fashion. Treatment-emergent thrombocytosis as a laboratory-based abnormality will be defined as an increase in platelet count from a maximum baseline value ≤600 billion/L to any postbaseline value >600 billion/L (Lengfelder et al. 1998). Planned and unplanned measurements will be included.

A listing of patients with treatment-emergent thrombocytosis may be provided for safety review.

6.15.5.3. Lipids Effects

Lipid effects will be assessed through analysis of elevated total cholesterol, elevated LDL cholesterol, decreased HDL cholesterol, and elevated triglycerides as described in Section 6.15.3 and with TEAEs potentially related to hyperlipidemia.

Categorical analyses will be performed using National Cholesterol Education Program (NCEP) Adult Treatment Panel (ATP) III guidelines (2002) as shown in the compound level safety standards. The grade-like categories shown in this table are ordered from traditionally most desirable to least desirable for the purposes of these analyses.

Shift tables will show the number and percentage of patients based on baseline to the least desirable category during the treatment period, with baseline depicted by the least desirable category during the baseline period. With each shift table, a shift table summary displaying the number and percentage of patients with the least desirable postbaseline results will be presented by treatment group for each treatment period within the following categories:

- decreased: postbaseline category more desirable than baseline category,
- increased: postbaseline category less desirable than baseline category,
- same: postbaseline category = baseline category.

Treatment-emergent laboratory abnormalities related to elevated total cholesterol, elevated triglycerides, elevated LDL cholesterol, and decreased and increased HDL cholesterol occurring at any time during the treatment period will be tabulated using the NCEP categories shown in the compound level safety standards.

Treatment-emergent elevated total cholesterol will be characterized as follows:

- increase to categories ‘Borderline high’ or ‘High’
- increase to category ‘High.’

Treatment-emergent elevated triglycerides will be characterized as

- increase to categories ‘Borderline high,’ ‘High,’ or ‘Very high’
- increase to categories ‘High’ or ‘Very high’
- increase to category ‘Very high.’

Treatment-emergent elevated LDL cholesterol will be characterized as

- increase to categories ‘Borderline high,’ ‘High,’ or ‘Very high’

- increase to categories ‘High’ or ‘Very high’
- increase to ‘Very high’

Treatment-emergent abnormal HDL cholesterol will be characterized as

- decreased HDL
  - decrease to categories ‘Normal’ or ‘Low’
  - decrease to category ‘Low’
- increased HDL
  - increase to categories ‘Normal’ or ‘High’
  - increase to category ‘High’

The percentages of patients with treatment-emergent potential hyperlipidemia will be summarize by treatment, ordered by decreasing frequency in the baricitinib 4-mg group using a predefined MedDRA list of PTs that is a subset of the narrow scope PTs in the MedDRA SMQ ‘Dyslipidemia’ (code 200000026) [see Compound level safety standards].

6.15.5.4. Renal Function Effects

Effects on renal function will be assessed through analysis of elevated creatinine.

CTCAEs will be applied for laboratory tests related to renal effects as shown in the compound level safety standards. This CTCAE grading scheme is consistent with both Version 3.0 and Version 4.03 of the CTCAE guidelines. Shift tables will show the number and percentage of patients based on baseline to maximum during the treatment period, with baseline depicted by highest grade during the baseline period. Treatment-emergent laboratory abnormalities related to elevated creatinine occurring at any time during the analysis period will be tabulated. Refer to the Compound level safety standards for details.

6.15.5.5. Elevations in Creatine Phosphokinase (CPK)

Elevations in creatine phosphokinase (CPK) will be addressed using CTCAE criteria as described in the compound level safety standards. This CTCAE grading scheme is consistent with both Version 3.0 and Version 4.03 of the CTCAE guidelines. Analyses will be the same as the CTCAE analyses specified for laboratory tests related to renal function events in Section 6.15.5.2.

A listing of elevated CPK (CTCAE grade of 3 or above) will be provided for medical safety review.

Treatment-emergent adverse events potentially related to muscle symptoms may be analyzed, based on reported AEs. The Muscle Symptoms special search category is a pre-defined MedDRA search criteria list that contains the narrow scope terms from the Rhabdomyolysis / myopathy SMQ (code 20000002) plus selected terms from the Musculoskeletal SOC. These terms are shown in the compound level safety standards.

6.15.5.6. Infections

Infections will be defined using all the PTs from the Infections and Infestations SOC as defined in MedDRA. Serious infection will be defined as all the infections that meet the SAE criteria.

The number and percentage of patients with TEAEs of infections, serious infections, and infections resulting in permanent study drug discontinuation will be summarized by treatment group using MedDRA PTs. The proportion of patients developing skin infections requiring antibiotic treatment by Week 16 will also be summarized on the overview of infections table.

The number and percentage of patients with TEAEs of infections by maximum severity will be summarized by treatment group using MedDRA PTs.

Treatment-emergent infections will be reviewed in context of other clinical and laboratory parameters via a listing (details see Compound level safety standards).

The TEAE infections will be further analyzed in terms of potential opportunistic infection, herpes zoster and herpes simplex. Summary of HBV DNA monitoring results and association between infection and neutropenia/lymphopenia will also be provided in the context of infections.

Opportunistic infection

To identify potential opportunistic infections (POIs), the following approach will be used:

- identifying the POIs using a list of MedDRA PTs (refer to the compound level safety standards).

Potential opportunistic infections identified through these search approaches may be combined in one list for medical assessment and final classification of whether the case met the modified Winthrop definitions for opportunistic infections (OIs).

A final listing for OIs will be provided for the CSR and to assist the composition of patient narratives.

Herpes zoster

Cases of herpes zoster will be further classified as follows:

- localized or non-multidermatomal-involvement of the primary and/or adjacent dermatomes only
    - complicated – documented ocular (cornea or deeper structure; for example, iritis, keratitis, retinitis, etc.) or motor nerve involvement (e.g., palsy; post herpetic neuralgia [PHN] does not meet criteria for motor nerve involvement).
    - uncomplicated-localized or non-multidermatomal cases that are not complicated
- multidermatomal-involvement beyond primary and adjacent dermatomes (that is, >3 contiguous dermatomes) or involvement of two or more non-contiguous dermatomes
    - complicated-documented ocular (cornea or deeper structure; for example iritis, keratitis, retinitis, etc.) or motor nerve involvement
    - uncomplicated-multidermatomal cases
- disseminated-systemic infection, visceral or widespread cutaneous (e.g., ≥5 dermatomes or 3 to 4 dermatomes including at least 1 non-contiguous).

- Recurrent - >1 infection occurring in an individual patient during the course of participation in the baricitinib clinical program.

All herpes zoster will undergo medical review to determine the classification as described above.

A summary of herpes zoster table will be provided. The summary table will also include event maximum severity, seriousness, whether resulting in temporary study drug interruption, whether resulting in study drug discontinuation, whether treated with antiviral medication, and event outcome. The incidence rate adjusted for observation time will also be provided (as defined in Section 6.15.2). Of note, in the context of herpes zoster, antiviral medication treatment is defined as that the medication was initiated at the event start date, or within 30 days before or after the event start date. The antiviral medication for herpes zoster includes but not limited to Aciclovir, Brivudine, Cidofovir, Famciclovir, Foscarnet, Ganciclovir, Penciclovir, Valaciclovir, Valganciclovir, Vidarabine (best presented by J05AB, J05AC, J05AE, and J05AH Anatomical Therapeutic Classification codes). Medical representatives will review the concomitant medication list prior to database lock and make adjustment of the above list if necessary.

If a patient has more than 1 event of herpes zoster, the event with the maximum severity will be used in these summary tables. If more than 1 event of herpes zoster occurs with the same severity, the event with the longest duration will be used in the summary table.

Herpes simplex

A summary analysis of herpes simplex will be provided. Herpes simplex will be defined based on MedDRA PT as listed in the compound level safety standards (both narrow and broad terms in the herpes simplex section). The list needs to be reviewed by medical prior to data locks (final and interim). The summary table will include event maximum severity, seriousness, whether resulting in temporary study drug interruption, whether resulting in study drug discontinuation, and whether treated with antiviral medication.

If a patient has more than 1 event of herpes simplex, the event with the maximum severity will be used in these summary tables. If more than 1 event of herpes simplex occurs with the same severity, the event with the longest duration will be used in the summary table.

Skin Infections

A summary analysis of skin infections may be provided. Skin infections may be defined based on MedDRA preferred term (see the Compound level safety standards).

HBV DNA

A listing of patients with detectable hepatitis B virus (HBV) deoxyribonucleic acid (DNA) post baseline will be provided.

HBV DNA status post baseline (not detectable, detectable but not quantifiable [i.e., < lower limit of detection (LLOD)], quantifiable [i.e., ≥LLOD]) will be summarized by treatment group stratified by baseline HBV serology status, specifically:

- HBsAb+ / HBcAb+

- HBsAb- / HBcAb+

Association between infection and neutropenia/lymphopenia

Depending on the number of cases with CTCAE Grade 2 or greater, a summary table will be provided for treatment-emergent infections that were preceded or accompanied by neutropenia. For this analysis, neutropenia is defined as (1) CTCAE Grade 2 or greater, (2) Grade 3 or greater. Infection events with onset date ≤14 days before or after the Grade 2 or greater neutrophil count collection date will be considered as infections preceded or accompanied by neutropenia.

Similar analyses as above will be conducted to evaluate the association between infection and lymphopenia.

6.15.5.7. Major Adverse Cardiovascular Events (MACE) and Other Cardiovascular Events

Potential major adverse cardiovascular events (MACE) and other cardiovascular events requiring adjudication will be analyzed.

Categories and subcategories analyzed will include, but are not limited to, the following:

- MACE
  - Cardiovascular death,
  - Myocardial infarction (MI),
  - Stroke,
- Other cardiovascular events
  - Transient ischemic attack,
  - Hospitalization for unstable angina,
  - Hospitalization for heart failure,
  - Serious arrhythmia,
  - Resuscitated sudden death,
  - Cardiogenic shock,
  - Coronary interventions (such as coronary artery bypass surgery or percutaneous coronary intervention),
- Non-cardiovascular death,
- All-cause death.

In general, events requiring adjudication are documented by investigative sites using an endpoint reporting CRF. This CRF is then sent to the adjudication center for external adjudication which uses an adjudication reporting CRF to document the final assessment of the event as a MACE, as some other cardiovascular event, or as no event (according to the Clinical Endpoint Committee Charter). In some cases, however, the investigator may not have deemed that an event had met the endpoint criteria but the event was still sent for adjudication as a potential MACE, other cardiovascular event, or no event. These events are included in the adjudication process to ensure adequate sensitivity. In these instances, the adjudication reporting CRF will not have a matching endpoint reporting CRF from the investigator. Events generated from these

circumstances will be considered as events sent for adjudication in the absence of an investigator's endpoint reporting form.

The number and percentage of patients with MACE, other cardiovascular events, non-cardiovascular death, and all-cause death, as positively adjudicated, will be summarized by treatment group based on the categories and subcategories above.

A listing of the events sent for adjudication will be provided to include data concerning the MedDRA PT related to the event, the seriousness of the event, and the event outcome, along with the adjudicated result.

6.15.5.8. Venous and Pulmonary Artery Thromboembolic (VTE) Events

Events identified as representative of venous thromboembolic events (VTE) disease will be classified as Deep Vein Thrombosis (DVT), pulmonary embolism (PE), or other peripheral venous thrombosis and will be analyzed. The following definitions apply:

- DVT: Clinical diagnosis of a thrombosis in a deep vein above the knee that must be confirmed by objective evidence of either: a filling defect of deep veins of the leg on venography or a non-compressible venous segment on ultrasound or confirmation by other imaging modality (e.g., Computed tomography [CT], Magnetic Resonance Imaging [MRI]).
- PE: Clinical diagnosis of pulmonary embolus that must be confirmed by objective evidence of either: a filling defect of pulmonary arteries by either pulmonary angiography or CT angiography or by a high probability Ventilation Perfusion (VQ) scan.
- Other Peripheral Venous Thrombosis: Clinical diagnosis of a venous thrombosis not specified by either DVT or PE above. Other peripheral venous thrombosis disease must be confirmed by objective evidence by imaging including venography, ultrasound, CT scan, or MRI. Examples of these would include non-superficial below knee thrombosis, portal vein, subclavian vein, or mesenteric vein. Superficial thrombophlebitis alone is not considered a VTE event.

In general, events requiring adjudication are documented by investigative sites using an endpoint reporting CRF. Refer to Section 6.15.5.7 for more details as the process is the same as that of MACE.

The number and percentage of patients with a VTE, DVT/PE, DVT, PE, and other peripheral venous thrombosis, as positively adjudicated, will be summarized by treatment group. Note that the below knee thrombosis captured in the "other peripheral venous thrombosis" category will be summarized within DVT.

A listing of the VTE events sent for external adjudication will be provided to include data concerning the MedDRA PT related to the event, the seriousness of the event, and the event outcome, along with the adjudicated result.

6.15.5.9. Arterial Thromboembolic (ATE) Events

Refer to the Compound level safety standards.

6.15.5.10. Malignancies

Malignancies will be identified using terms from the malignant tumors SMQ (SMQ 20000194). Malignancies excluding non-melanoma skin cancers (NMSC) and NMSC will be reported separately.

All the cases identified by malignant tumors SMQ will be assessed through medical review to determine confirmed NMSC cases.

First, a listing including all the malignancy cases will be prepared before database lock along with the *planned* NMSC flag according to the current MedDRA version PTs (the list will be updated depending on the MedDRA version used for analysis):

- Squamous cell carcinoma of skin (10041834)
- Bowen's disease (10006059)
- Basal cell carcinoma (10004146)
- Basosquamous carcinoma (10004178)
- Basosquamous carcinoma of skin (10004179)
- Squamous cell carcinoma (10041823)
- Skin squamous cell carcinoma metastatic (10077314)
- Skin cancer (10040808)
- Carcinoma in situ of skin (10007390)
- Keratoacanthoma (10023347)
- Vulvar squamous cell hyperplasia (10079905)
- Skin squamous cell carcinoma recurrent (10081136)

This internal review is to occur prior to database lock. The case review and subsequent summary analyses will include all the cases reported in the study database or by LSS report, disregarding the length of gap between the last treatment dose date and the event date. The NMSC flag will be confirmed during the internal review process.

The number and percentage of patients with TEAEs associated malignancies excluding NMSC and NMSC will be summarized by treatment group.

6.15.5.11. Allergic Reactions/Hypersensitivities

A search will be performed using the MedDRA version 21.1 SMQs to search for relevant events, using the following queries:

- Anaphylactic reaction SMQ (20000021)
- Hypersensitivity SMQ (20000214)
- Angioedema SMQ (20000024)

Assessment of the Anaphylactic reaction SMQ includes an algorithmic query. The algorithmic approach comprises one or more events associated with an individual administration of study drug, where the events include:

- A narrow term from the SMQ (Category A of the SMQ);
- Multiple terms from the SMQ, comprising terms from at least two of the following categories from the SMQ:
    - Category B - (Upper Airway/Respiratory signs and symptoms)
    - Category C - (Angioedema/Urticaria/Pruritus/Flush signs and symptoms)
    - Category D - (Cardiovascular/Hypotension signs and symptoms).

Refer to the Compound safety level standards for details.

6.15.5.12. Gastrointestinal Perforations

Treatment-emergent adverse events related to potential gastrointestinal (GI) perforations will be analyzed using reported AEs. Identification of these events will be based on review of the PTs of the MedDRA SMQ 20000107, GI perforations (note that this SMQ holds only narrow terms and has no broad terms). Potential GI perforations identified by the above SMQ search will be provided as a listing for internal review. Each case will be assessed to determine whether it is GI perforation. A summary table based on medical review may be provided and treatment comparisons will be made using Fisher's exact test.

6.15.5.13. Columbia Suicide Severity Rating Scale

Suicidal ideation, suicidal behavior, and self-injurious behavior without suicidal intent, based on the C-SSRS, will be listed by patient and visit. Only patients that show suicidal ideation/behavior or self-injurious behavior without suicidal intent during treatment will be displayed along with all their ideation and behavior, even if not positive (i.e., if a patient's answers are all 'no' for the C-SSRS, then that patient will not be displayed). A summary of the C-SSRS categories during treatment and a shift summary in the C-SSRS categories from baseline during treatment will be provided. Refer to the Compound safety level standards for details.

6.15.5.14. Self-Harm Supplement Form and Self-Harm Follow-up Form

The Self-Harm Supplement Form is a single question to enter the number of suicidal behavior events, possible suicide behaviors, or nonsuicidal self-injurious behaviors. If the number of behavioral events is greater than zero, it will lead to the completion of the Self-Harm Follow-Up Form. The Self-Harm Follow-Up Form is a series of questions that provides a more detailed description of the behavior cases. A listing of the responses give on the Self-Harm Follow-Up Form will be provided.

6.16. Subgroup Analyses

Subgroup analyses comparing each dose of baricitinib to placebo will be performed on the ITT population at Week 16 using the primary censoring rule for the following:

- Proportion of patients achieving IGA 0 or 1
- Proportion of patients achieving EASI75 Response Rate
- Proportion of patients achieving Itch NRS 4-point improvement

The following subgroups, categorized into disease-related characteristics and demographic characteristics, will be evaluated:

- Patient Demographic and Characteristics Subgroups:
  - Gender (male, female)
  - Age group (<65, ≥65 years old)
  - Age group (<65, ≥65 to <75, ≥75 to <85, ≥85 years old)
  - Baseline weight: (<60 kg, ≥60 to <100 kg, ≥100 kg)
  - Baseline BMI (<25 kg/m^2, ≥25 to <30 kg/m^2, ≥30 kg/m^2)
  - Race: (American Indian or Alaska Native, Asian, Black or African American, Native Hawaiian or Other Pacific Islander, White, Multiple)
  - Baseline renal function status: impaired (eGFR <60 mL/min/1.73 m^2) or not impaired (eGFR ≥60 mL/min/1.73 m^2)
- Geographic Region Subgroups:
  - Region: (as defined in Table JAIY.5.1)
  - Specific regions (Europe, other)
  - Specific regions (East Asia[Korea, Japan and Taiwan], other)
  - Specific country (Japan, other)
  - Prior systemic therapy use (yes, no)
- Baseline Disease-Related Characteristics Subgroup
  - Baseline disease severity (IGA score): 3, 4

Descriptive statistics will be provided for each treatment and stratum of a subgroup as outlined, regardless of sample size. As all endpoints are categorical, subgroup analyses will be performed using logistic regression using Firth's correction to accommodate (potential) sparse response rates. The model will include the categorical outcome as the dependent variable and baseline value (for EASI and itch), baseline severity, treatment, subgroup, and treatment-by-subgroup interaction as explanatory variables. Missing data will be imputed using NRI (Section 6.4.1). The treatment-by-subgroup interaction comparing treatment groups will be tested at the 0.1 significance level. The p-value from the logistic regression model will be reported for the interaction test and the subgroup test, unless the model did not converge. Response counts and percentages will be summarized by treatment for each subgroup category. The difference in percentages and 95% CI of the difference in percentages using the Newcombe-Wilson without continuity correction will be reported. The corresponding p-value from the Fisher's exact test will also be produced.

In case any level of a subgroup comprises <10% of the overall sample size, only descriptive summary statistics will be provided for treatment arms, and no treatment group comparisons will be performed within these subgroup levels.

Additional subgroup analyses on efficacy may be performed as deemed appropriate and necessary.

6.17. Protocol Deviations

Protocol deviations will be tracked by the clinical team, and their importance will be assessed by key team members during protocol deviation review meetings. Out of all important protocol deviations (IPDs) identified, a subset occurring during Period 2 with the potential to affect efficacy analyses will result in exclusion from the PP population.

Potential examples of deviations include patients who receive excluded concomitant therapy, significant non-compliance with study medication (<80% of assigned doses taken, failure to take study medication and taking incorrect study medication), patients incorrectly enrolled in the study, and patients whose data are questionable due to significant site quality or compliance issues. Refer to a separate document for the important protocol deviations.

Trial Issue Management Plan includes the categories and subcategories of important protocol deviations and whether or not these deviations will result in the exclusion of patients from per protocol set.

The number and percentage of patients having IPD(s) will be summarized within category and subcategory of deviation by treatment group for Period 2 using the ITT population. Individual patient listings of IPDs will be provided. A summary of reasons patients were excluded from the PPS will be provided by treatment group.

6.18. Interim Analyses and Data Monitoring

An interim analyses may be conducted at the time when the last patient completes Visit 8 (Week 16) or ETV.

The baricitinib AD, AA and SLE Phase 3 programs Data Monitoring Committee (DMC) is an independent expert advisory group commissioned and charged with the responsibility of evaluating cumulative safety at regular intervals. As such, the primary objective of the DMC is to monitor the safety of the subjects enrolled in the baricitinib AD, AA and SLE Phase 3 programs by reviewing the available clinical data at scheduled time points, as described in this DMC Charter, as well as on an ad hoc basis, as needed. The DMC will consist of members external to Lilly. This DMC will follow the rules defined in the DMC charter, focusing on potential and identified risks for this molecule and for this class of compounds. Data Monitoring Committee membership will include, at a minimum, specialists with expertise in dermatology, statistics, cardiology, and other appropriate specialties.

The DMC will be authorized to review unblinded results of analyses by treatment group prior to database lock, including study discontinuation data, AEs including SAEs, clinical laboratory data, vital sign data, etc. The DMC may recommend continuation of the study, as designed; temporary suspension of enrollment; or the discontinuation of a particular dose regimen or the entire study. While the DMC may request to review efficacy data to investigate the benefit/risk relationship in the context of safety observations for ongoing patients in the study, no information regarding efficacy will be communicated. Moreover, the study will not be stopped for positive efficacy results nor will it be stopped for futility. Hence, no alpha is spent. Details of the DMC, including its operating characteristics, are documented in the Baricitinib Atopic

Dermatitis DMC charter and further details are given in the Interim Analysis Plan in Section 6.18.1.

Besides DMC members, a limited number of pre-identified individuals may gain access to the limited unblinded data, as specified in the unblinding plan, prior to the interim or final database lock, for preparation of regulatory documents. Information that may unblind the study during the analyses will not be reported to study sites or blinded study team until the study has been unblinded.

6.18.1. Interim Analysis Plan

The final analyses for this study are considered as the analyses to be presented to the DMC.

6.19. Planned Exploratory Analyses

The planned exploratory analyses are described in Sections 6.12 and 6.13. Additional exploratory analyses may be conducted such as exploring inadequate or super responders and their baseline characteristics and will be documented in a supplemental SAP. Health Technology Assessment (HTA) toolkit analyses, which may be produced, will also be documented in the supplemental SAP.

6.20. Annual Report Analyses

Annual report analyses, such as the Development Update Safety Report (DSUR), will be documented in a separate document.

6.21. Clinical Trial Registry Analyses

Additional analyses will be performed for the purpose of fulfilling the Clinical Trial Registry (CTR) requirements.

Analyses provided for the CTR requirements include a summary of AEs, provided as a dataset which will be converted to an XML file. Both SAEs and 'Other' AE are summarized: by treatment group, by MedDRA PT.

- An AE is considered 'Serious' whether or not it is a TEAE.
- An AE is considered in the 'Other' category if it is both a TEAE and is not serious. For each SAE and 'Other' AE, for each term and treatment group, the following are provided:
  - the number of participants at risk of an event
  - the number of participants who experienced each event term
  - the number of events experienced.
- Consistent with www.ClinicalTrials.gov requirements, 'Other' AEs that occur in fewer than 5% of patients/subjects in every treatment group may not be included if a 5% threshold is chosen (5% is the minimum threshold).
- AE reporting is consistent with other document disclosures for example, the CSR, manuscripts, and so forth.

Similar methods will be used to satisfy the European Clinical Trials Database (EudraCT) requirements.

7. Unblinding Plan

A separate JAIY Blinding / Unblinding Plan contains details of how the blind is maintained for this study.

8. References

Alosh M, Bretz F, Huque M. Advanced multiplicity adjustment methods in clinical trials. *Stat Med.* 2014;33(4):693-713.

Basra MK, Salek MS, Camilleri L, Sturkey R, Finlay AY. Determining the minimal clinically important difference and responsiveness of the Dermatology Life Quality Index (DLQI): further data. *Dermatology*. 2015;230(1):27-33.

Bretz F, Posch M, Glimm E, Klinglmueller F, Maurer W, Rohmeyer K. Graphical approaches for multiple comparison procedures using weighted Bonferroni, Simes, or parametric tests. *Biom J*. 2011;53(6):894-913.

Charman CR, Venn AJ, Williams, HC. The Patient-Oriented Eczema Measure: development and initial validation of a new tool for measuring atopic eczema severity from the patients' perspective. *Arch Dermatol.* 2004;140:1513-1519.

[CTCAE] Common Terminology Criteria for Adverse Events, Cancer Therapy Evaluation Program, Version 3.0, DCTD, NCI, NIH, DHHS, March 31, 2003. Available at: https://ctep.cancer.gov/protocoldevelopment/electronic_applications/docs/ctcaev3.pdf. Accessed 25 July 2017.

[CTCAE] Common Terminology Criteria for Adverse Events, Version 4.03, DHHS NIH NCI, NIH Publication No. 09-5410, June 14, 2010. Available at: https://www.eortc.be/services/doc/ctc/CTCAE_4.03_2010-06-14_QuickReference_5x7.pdf. Accessed 25 July 2017.

EuroQol Group. EQ-5D-5L User Guide. Version 2.1. Available at: http://www.euroqol.org/fileadmin/user_upload/Documenten/PDF/Folders_Flyers/EQ-5D-5L_UserGuide_2015.pdf. Accessed March 2018.

Gillings D, Koch G. The application of the principle of intention-to-treat to the analysis of clinical trials. *Drug Information Journal.* 1991;25(3):411-424.

Hanifin JM, Thurston M, Omoto M, Cherill R, Tofte SJ, Graeber M. The eczema area and severity index (EASI): assessment of reliability in atopic dermatitis. EASI Evaluator Group. *Exp Dermatol*. 2001;10(1):11-18.

Herdman M, Gudex C, Lloyd A, Janssen MF, Kind P, Parkin D, Bonsel G, Badia X. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). *Qual Life Res*. 2011;20(10):1727-1736.

Hongbo Y, Thomas CL, Harrison MA, Salek MS, Finlay AY. Translating the science of quality of life into practice: What do dermatology life quality index scores mean? *J Invest Dermatol*. 2005;125(4):659-664.

[ICH] International Conference on Harmonisation. Harmonised Tripartite Guideline: Clinical safety data management: definitions and standards for expedited reporting. E2A. 1994;Step 4. Available at: http://www.ich.org/products/guidelines/efficacy/article/efficacy-guidelines.html. Accessed 21 July 2017.

[ICH E9 R1] Estimands and Sensitivity Analysis in Clinical Trials http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E9/E9-R1EWG_Step2_Guideline_2017_0616.pdf

Khilji FA, Gonzalez M, Finlay AY. Clinical meaning of change in Dermatology Life Quality Index scores. *Br J Dermatol*. 2002;147(suppl 62):25-54.

Kimball AB, Naegeli AN, Edson-Heredia E, Lin CY, Gaich C, Nikai E, Yosipovitch G. Psychometric properties of the Itch Numeric Rating Scale in patients with moderate-to-severe plaque psoriasis. *Br J Dermatol.* 2016;175(1):157-162.

Kunz B, Oranje A, Labrèze, Stadler JF, Ring J, Taïeb A, Clinical Validation and Guidelines for the SCORAD Index: Consensus Report of the European Task Force Atopic Dermatitis. *Dermatology*. 1997;195(1):10-19.

Lengfelder E, Hochhaus A, Kronawitter U, Hoche D, Queisser W, Jahn-Eder M, Burkhardt R, Reiter A, Ansari H, Hehlmann R. Should a platelet limit of 600 x 10(9)/l be used as a diagnostic criterion in essential thrombocythaemia? An analysis of the natural course including early stages. *Br J Haematol*. 1998;100(1):15-23.

Naegeli AN, Flood E, Tucker J, Devlen J, Edson-Heredia E. The Worst Itch Numeric Rating Scale for patients with moderate to severe plaque psoriasis or psoriatic arthritis. *Int J Dermatol.* 2015;54(6):715-722.

[NCEP] National Cholesterol Education Program. Third Report of the NCEP Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel [ATP] III) Final Report, National Cholesterol Education Program, National Heart, Lung, and Blood Institute, National Institutes of Health, NIH Publication No. 02-5215, September 2002. *Circulation*. 2002;106(25):3143-421.

Program Safety Analysis Plan: Baricitinib (LY3009104) Version 5. Report on file, Eli Lilly and Company.

Protocol I4V-MC-JAIY(a). A Multicenter, Randomized, Double-Blind, Placebo-Controlled, Phase 3 Study to Evaluate the Efficacy and Safety of Baricitinib in Adult Patients with Moderate to Severe Atopic Dermatitis. Report on file, Eli Lilly and Company.

Reilly MC, Zbrozek AS, Dukes EM. The validity and reproducibility of a work productivity and activity impairment instrument. *Pharmacoeconomics*. 1993;4(5):353-365.

Schram ME, Spuls PI, Leeflang MMG, Lindeboom R, Bos JD, Schmitt J. EASI, (objective) SCORAD and POEM for atopic eczema: responsiveness and minimal clinically important difference. *Allergy.* 2012;67(1):99-106.

Snaith RP. The Hospital Anxiety and Depression Scale. *Health and Quality of Life Outcomes*. 2003;1:29.

Stalder J, Taieb A, Atherton D, Bieber P, Bonifazi E, Broberg A, Calza A, Coleman R, De Prost Y, Stalder J, Gelmetti C, Cuannetti A, Harper J, Kunz B, Lachapelle J, Langeland T, Lever R, Oranje A, Oueille-Roussel C, Revuz J, Ring J, Roujeau J, Saurat J, Song M, Tennstedt D, Van Neste D, Vieluf D, Poncet M. Severity scoring of atopic dermatitis: the SCORAD index. Consensus Report of the European Task Force on Atopic Dermatitis. *Dermatology*. 1993;186(1):23-31.

US National Archives and Records Administration. Code of Federal Regulations (CFR). Investigational New Drug Application (IND) safety reporting. 2010;Title 21:Section 312.32. Available at: https://www.ecfr.gov/cgi-bin/text-idx?SID=fd87d5f6fb00b95672dbe924a6d3b2f0&mc=true&node=se21.5.312_132&rgn=div8. Accessed 21 July 2017.

White D, Leach C, Sims R, Atkinson M, Cottrell D. Validation of the Hospital Anxiety and Depression Scale for use with adolescents. *Br J Psychiatry*. 1999;175(5):452-454.

Winthrop KL, Novosad SA, Baddley JW, Calabrese L, Chiller T, Polgreen P, Bartalesi F, Lipman M, Mariette X, Lortholary O, Weinblatt ME, Saag M, Smolen J. Opportunistic infections and biologic therapies in immune-mediated inflammatory diseases: consensus recommendations for infection reporting during clinical trials and postmarketing surveillance. *Ann Rheum Dis*. 2015;74(12):2107-2116.

Zigmond AS, Snaith RP. The Hospital Anxiety and Depression Scale. *Acta Psychiatr Scand*. 1983;67(6):361-370.

1. Statistical Analysis Plan for EU: I4V-MC-JAIY (a): A Multicenter, Randomized, Double-Blind, Placebo-Controlled, Phase 3 Study to Evaluate the Efficacy and Safety of Baricitinib in Combination with Topical Corticosteroids in Adult Patients with Moderate to Severe Atopic Dermatitis

BREEZE-AD7

Baricitinib (LY3009104) Atopic Dermatitis

Study I4V-MC-JAIY is a Phase 3, multicenter, randomized, double-blind, placebo-controlled, parallel-group, outpatient study evaluating the efficacy and safety of baricitinib 4 mg once daily (QD) plus topical corticosteroids (TCS) and 2 mg QD plus TCS, as compared to placebo plus TCS in adult patients with moderate to severe atopic dermatitis.

Eli Lilly and Company
Indianapolis, Indiana USA 46285
Protocol I4V-MC-JAIY
Phase 3
Statistical Analysis Plan Version 1 electronically signed and approved by Lilly:
01 April 01, 2019
Statistical Analysis Plan Version 2 electronically signed and approved by Lilly on date provided below.

Approval Date: 12-Aug-2019 GMT

2. Table of Contents

Section **Page**

Table of Contents

Table **Page**

Table of Contents

Figure **Page**

3. Revision History

Statistical Analysis Plan (SAP) Version 1 is based on Protocol I4V-MC-JAIY(a). Statistical Analysis Plan Version 2 was approved prior to unblinding. A summary of changes between Version 1 and Version 2 are as follows:

Section	Summary of Changes
4.3. Exploratory objectives	• Updated the time frame for the objective of frequency of patient-reported "no itch" and "no pain" by starting from Week 0. • Added new exploratory analyses on Hospital Anxiety Depression (HADS), Dermatology Life Quality Index (DLQI), Atopic Dermatitis Sleep Scale (ADSS), and Patient-Oriented Eczema Measure (POEM). • Added new exploratory analyses on early responders.
6.11.1. Background TCS	• Updated the weight of dispensed topical corticosteroids (TCS) tube to align with JAHL SAP.
6.2.2. Definition on Baseline and Postbaseline Measures	• Updated the derivations of itch weekly scores.
6.2.3. Analysis Methods	• Updated the structures of covariance matrix in mixed model repeated measures (MMRM) model.
6.6. Multiple Comparisons	• Updated the graphical testing procedure.
6.16 Subgroup Analysis	• Added subgroup analyses for East Asia vs. other.
Table 6.6	• Updated the endpoints on the number days of itch-free and pain-free days.
Table 6.7	• Updated analyses on the number of itch-free and pain-free days. • Change the type of analysis of covariance (ANCOVA) (modified last observation carried forward [mLOCF]) analyses on PROMIS from "exploratory analyses" to "sensitivity analyses."

4. Study Objectives

4.1. Primary Objective

The primary objective of this study is to test the hypothesis that baricitinib 4 mg once daily (QD) plus topical corticosteroids (TCS) or baricitinib 2 mg QD plus TCS is superior to placebo plus TCS in the treatment of patients with moderate to severe atopic dermatitis (AD), as assessed by the proportion of patients achieving the validated Investigator's Global Assessment for AD (vIGA-AD, referred to throughout the SAP as IGA) of 0 or 1 with a ≥2-point improvement at Week 16.

In particular, the associated estimand for this objective is to measure the effect of therapy with baricitinib as assessed by the proportion of patients with a response of IGA 0 or 1 at Week 16 assuming treatment response disappears after patients are rescued or discontinue from the study or treatment. See Sections 6.4.1 and 6.12.1 on how this estimand handles outcomes after the occurrence of any intercurrent event through nonresponder imputation (NRI).

4.2. Secondary Objectives

4.2.1. Key Secondary Objectives

These are prespecified objectives that will be adjusted for multiplicity.

Objectives	Endpoints
To compare the efficacy of baricitinib 2 mg QD + TCS or baricitinib 4 mg QD + TCS to placebo + TCS in AD during the 16-week double-blind placebo-controlled treatment period as measured by improvement in signs and symptoms of AD.	• Proportion of patients achieving EASI75 at 16 weeks • Proportion of patients achieving EASI90 at 16 weeks • Percent change from baseline in EASI score at 16 weeks • Proportion of patients achieving SCORAD75 at 16 weeks
To compare the efficacy of baricitinib 2 mg QD + TCS or baricitinib 4 mg QD + TCS to placebo + TCS in AD during the 16-week double-blind placebo-controlled treatment period as assessed by patient-reported outcome measures.	• Proportion of patients achieving a 4-point improvement from baseline in Itch NRS at 2 days, 1 week, 2 weeks, 4 weeks, and 16 weeks • Mean change from baseline in the score of Item 2 of the ADSS at 1 week and 16 weeks • Mean change from baseline in Skin Pain NRS at 16 weeks

4.2.2. Other Secondary Objectives

These are prespecified objectives that will not be adjusted for multiplicity.

Objectives	Endpoints
To test the hypothesis that baricitinib 2 mg QD + TCS or baricitinib 4 mg QD + TCS is superior to placebo + TCS in the treatment of patients with moderate to severe AD.	• Proportion of patients achieving IGA of 0 or 1 with a ≥2-point improvement at Week 4
To compare the efficacy of baricitinib 2 mg QD + TCS or baricitinib 4 mg QD + TCS to placebo + TCS in AD during the 16-week double-blind placebo-controlled period as measured by signs and symptoms of AD.	• Proportion of patients achieving EASI50 at 16 weeks • Proportion of patients achieving IGA of 0 at 16 weeks • Mean change from baseline in SCORAD at 16 weeks • Proportion of patients achieving SCORAD90 at 16 weeks • Mean change from baseline in BSA affected at 16 weeks • Proportion of patients developing skin infections requiring antibiotic treatment by Week 16 • Mean gram quantity of background TCS used over 16 weeks (tube weights)
To compare the efficacy of baricitinib 2 mg QD + TCS or baricitinib 4 mg QD + TCS to placebo + TCS in AD during the 16-week, double-blind, placebo-controlled treatment period as assessed by patient-reported outcome/QoL measures.	• Percent change from baseline in Itch NRS at 2 days, 1 week, 4 weeks, and 16 weeks • Mean change from baseline in Itch NRS at 2 days, 1 week, 4 weeks and 16 weeks • Proportion of patients achieving a 4-point improvement from baseline in Skin Pain NRS at 16 weeks • Mean change from baseline in the total score of the POEM at 16 weeks • Mean change in the PGI-S-AD scores at 16 weeks • Mean change from baseline in HADS at 16 weeks • Mean change in the DLQI scores at 16 weeks • Mean change in the WPAI scores at 16 weeks • Mean change in the EQ-5D-5L scores at 16 weeks • Mean number of days without use of background TCS over 16 weeks

4.3. Exploratory Objectives

The exploratory objectives of this study are the following:

Objectives/Endpoints
• Frequency of patient-reported "no itch" (Itch NRS score = 0) days from daily diaries from Week 0 to Week 16 • Frequency of patient-reported "no pain" (Skin Pain NRS score = 0) days from daily diaries from Week 0 to Week 16 • Mean change from baseline in PIQ – Itch Interference score • Mean change from baseline in PIQ – Activity and Clothing score • Mean change from baseline in PIQ – Mood and Sleep score • Mean change from baseline in PIQ – Scratching Behavior score • Mean change from baseline in PROMIS – Sleep-Related Impairment score • Mean change from baseline in Neuro-QoL – Cognitive Function score • Patient Benefit Index score at 16 weeks – global score plus the following subscales: o Reducing social impairments o Reducing psychological impairments o Reducing impairments due to therapy o Reducing physical impairments o Having confidence in healing • Proportion of patients achieving PBI global score ≥1 at 16 weeks • Mean change from baseline in the score of Item 1 of the ADSS at 1 week and 16 weeks • Proportion of patients achieving a ≥1-point improvement in the score of Item 1 of the ADSS for those with baseline Item 1 score ≥1 • Proportion of patients achieving a ≥1-point improvement in the score of Item 2 of the ADSS for those with baseline Item 2 score ≥ 1 • Proportion of patients achieving a ≥2-point improvement in the score of Item 2 of the ADSS for those with baseline Item 2 score ≥2 • Mean change from baseline in the score of Item 3 of the ADSS at 1 week and 16 weeks • Proportion of patients achieving a ≥1-point improvement in the score of Item 3 of the ADSS for those with baseline Item 3 score ≥1 • To evaluate changes from baseline in immunoglobulin E levels during the study • To evaluate changes from baseline in eosinophil levels during the study • To assess time to 4-point Itch NRS improvement during the first 14 days after initiation of treatment • To assess time to 4-point improvement in Skin Pain during the first 14 days after the initiation of treatment • Proportion of patients achieving a ≥4-point improvement in DLQI total score for those with baseline DLQI total score ≥4 • Proportion of patients achieving DLQI total score 0 or 1 • Proportion of patients achieving a ≥ 4-point improvement in POEM total score for those with baseline total score ≥4 • Proportion of patients achieving HADS Anxiety Score <8 for those with baseline HADS Anxiety Score ≥8 • Proportion of patients achieving HADS Depression Score <8 for those with baseline HADS Depression Score ≥8 • Proportion of patients achieving improvement with HADS Anxiety Score or HADS Depression Score <8 for those with baseline HADS Anxiety Score ≥8 or HADS Depression Score ≥8 • Mean change from baseline in HADS subscale scores

- Proportion of patients achieving ≥4-point improvement in Itch NRS or IGA≤1 at Week 16 for those with ≥3-point improvement in Itch NRS or IGA≤2 at Week 4
- Proportion of patients achieving ≥ 4-point improvement in Itch NRS or IGA≤1 at any time between Week 8 and Week 16 for those with ≥3-point improvement in Itch NRS or IGA≤2 at Week 4
- Proportion of patients achieving [≥4-point improvement in Itch NRS and IGA≤2] or [IGA≤1] at Week 16 for those with ≥3-point improvement in Itch NRS or IGA≤2 at Week 4
- Proportion of patients achieving [≥4-point improvement in Itch NRS and IGA≤2] or [IGA≤1] at any time between Week 8 and Week 16 for those with ≥3-point improvement in Itch NRS or IGA≤2 at Week 4

5. Study Design

5.1. Summary of Study Design

Study I4V-MC-JAIY (JAIY) is a Phase 3, multicenter, randomized, double-blind, placebo-controlled, parallel-group, outpatient study evaluating the efficacy and safety of baricitinib 2 mg QD and 4 mg QD, in combination with TCS, as compared to placebo in combination with TCS, in adult patients with moderate to severe AD. The study is divided into 3 periods, a 5-week Screening period, a 16-week Double-Blinded Treatment period, and a 4-week Post-Treatment Follow-Up period. For those patients who complete the 16-week treatment period, there is an option to participate in the long-term extension study I4V-MC-JAHN (JAHN).

Approximately 300 patients ≥18 years of age who have responded inadequately to topical therapy will be randomized at a 1:1:1 ratio to receive placebo QD, baricitinib 2 mg QD, or baricitinib 4 mg QD in combination with TCS (100 patients in each treatment group). Patients will be stratified at randomization according to disease severity (Investigator's Global Assessment [IGA] 3 vs. 4) and geographic region.

Study JAIY will consist of 3 periods:

- Period 1: Screening period is between 8 and 35 days prior to Week 0 (Visit 2)
- Period 2: Double-Blind, Placebo-Controlled Treatment period from Week 0 (Visit 2) through Week 16 (Visit 8)
- Period 3: Post-Treatment Follow-Up period from last treatment visit at Week 16 (Visit 8) or Early Termination Visit (ETV) to approximately 28 days after the last dose of investigational product

Figure JAIY.5.1 illustrates the study design. The blinding procedure is described in the Protocol.

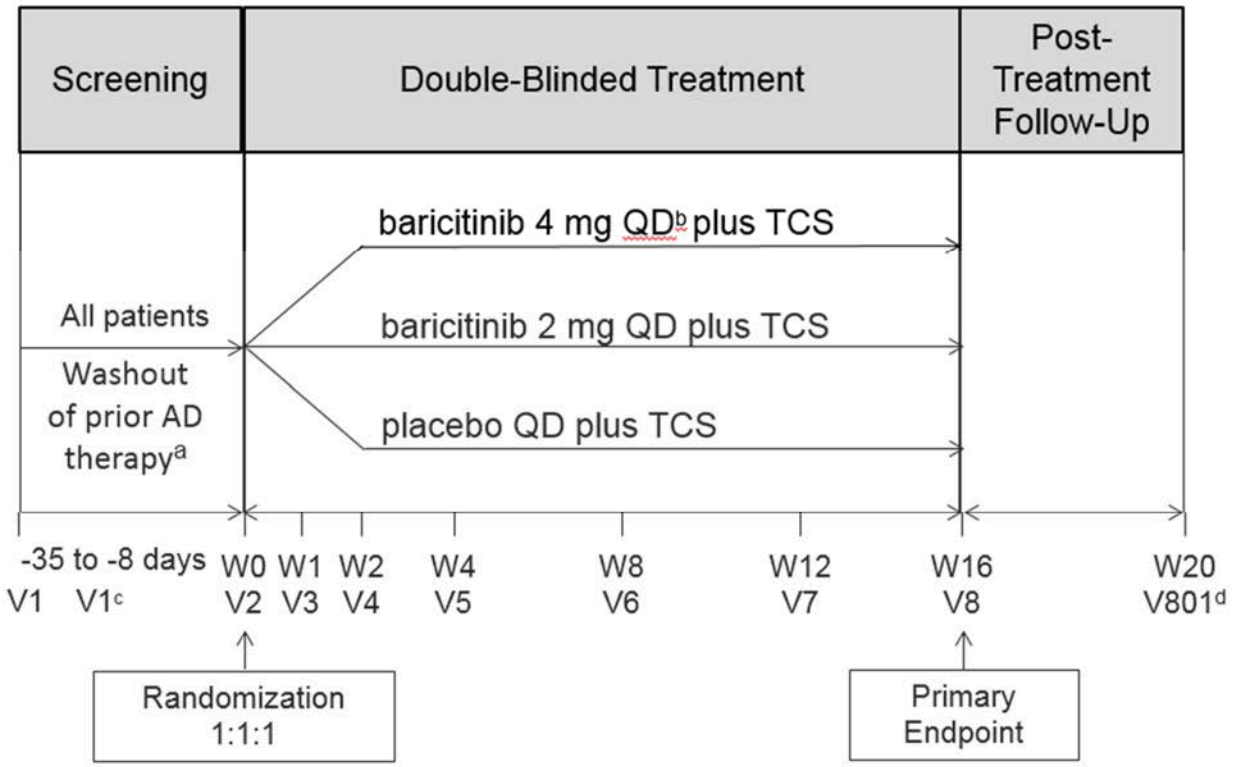


Figure JAIY.5.1. Illustration of study design for Clinical Protocol I4V-MC-JAIY.

Abbreviations: AD = atopic dermatitis; eGFR = estimated glomerular filtration rate; PPD = purified protein derivative; QD = once daily; TCS = topical corticosteroids; V = visit; W = week.

[a] Applicable to patients taking topical treatments (excluding emollients) or systemic treatments for AD at the time of screening.

[b] For patients randomized to the 4-mg once daily dose who have renal impairment (defined as eGFR <60 mL/min/1.73 m^2), the baricitinib dose will be 2 mg once daily.

[c] Patients for whom PPD skin test for the evaluation of tuberculosis infection was performed at V1 must return and PPD test must be read 48 to 72 hours after Visit 1 (post-PPD).

[d] Occurs approximately 28 days after the last dose of investigational product. Not required for those patients entering the long-term extension study JAHN.

5.2. Method of Assignment to Treatment

Patients who meet all criteria for enrollment will be randomized in a 1:1:1 ratio (placebo; baricitinib 2 mg; baricitinib 4 mg) to double-blind treatment at Visit 2 (Week 0). Randomization will be stratified by geographic region (Europe [EU], Japan [JPN], rest-of-world [ROW]) and disease severity at baseline (IGA 3 vs. 4). Assignment to treatment groups will be determined by a computer-generated random sequence using an IWRS. The IWRS will be used to assign packages containing double-blind investigational product tablets to each patient according to the study schedule of activities. Site personnel will confirm that they have located the correct packages by entering a confirmation number found on the packages into the IWRS.

This study will be conducted internationally in multiple sites. Table JAIY.5.1 describes how regions will be defined for stratification. Regions may be combined for statistical analyses in the

case when one of the region strata fails to meet the required minimum number of 30 patients. The 2 region strata with the least number of patients will then be pooled.

Table JAIY.5.1. Geographic Regions for Stratification

Region	Countries
Europe	Germany, Italy, Poland, Spain, Austria
Japan	Japan
Rest of World	Taiwan, Australia, Korea, Argentina

6. A Priori Statistical Methods

6.1. Determination of Sample Size

Study JAIY will aim to enroll approximately 300 patients ≥18 years of age. The proposed sample size will ensure a 89% power to detect an absolute difference of 20% between the baricitinib 4 mg and placebo treatment groups and the baricitinib 2 mg and placebo treatment groups, each using a 2-sided alpha of 0.025 and a Fisher´s exact test, assuming a 10% placebo response rate for the primary endpoint. The assumptions are based on what was observed in the Phase 2 study (JAHG). The proposed end point of IGA 0 or 1 represents patients whose AD is clear or almost clear from a baseline of moderate or severe disease. The anticipated effect size represents 2 times more patients achieving this benefit compared to placebo, which, in discussion with therapeutic experts, is of a magnitude that is considered clinically relevant.

Sample size and power estimates were obtained from nQuery® Advisor 7.0.

6.2. General Considerations

This plan describes *a priori* statistical analyses for efficacy, health outcomes, and safety that will be performed.

Statistical analysis of this study will be the responsibility of Eli Lilly and Company (Lilly). The statistical analyses will be performed using SAS® Version 9.4 or higher.

Not all displays described in this SAP will necessarily be included in the clinical study report (CSR). Not all displays will necessarily be created as a "static" display. Some may be incorporated into interactive display tools instead of or in addition to a static display. Any display described in this SAP and not included in the CSR will be available upon request.

Statistical tests of treatment effects and confidence intervals (CIs) will be performed at a 2-sided significance level of 0.05, unless otherwise stated (e.g., graphical multiple testing strategy in Section 6.6).

Data collected at ETVs will be mapped to the closest scheduled visit number for that patient if it falls within the visit window as discussed in Section 6.2.2. For by-visit summaries, only visits in which a measure was scheduled to be collected will be summarized. Any unscheduled visit data will be included at the patient-level listings. However, the data will still be used in other analyses, including shift analyses for safety analytes, change from baseline using modified last observation carried forward (mLOCF) for efficacy analyses, and other categorical analyses including safety.

6.2.1. Analysis Populations

Intent-to-treat (ITT) population: The ITT population analysis set is defined as all randomized patients.

Per-protocol Set (PPS): PP subset of the ITT analysis set will include those patients who do not have any significant or important protocol violations. Qualifications for and identification of

significant or important protocol violations will be determined while the study remains blinded, prior to database lock.

Follow-up population: The follow-up population is defined as patients who entered the follow-up period.

Unless otherwise specified, the efficacy and health outcome analyses will be conducted on the ITT population (Gillings and Koch 1991), which seeks to preserve the benefits of randomization and avoid selection bias. Patients will be analyzed according to the treatment to which they were randomized. In addition, the analyses of primary and key secondary endpoints will be repeated using the PPS population.

Safety population: The safety population is defined as all randomized patients who receive at least 1 dose of investigational product and who did not discontinue from the study for the reason 'Lost to Follow-up' at the first postbaseline visit.

Safety analyses will be performed using the safety population. Patients will be analyzed according to the treatment regimen to which they were assigned. Analyses of the safety endpoints, many of which are incidence based, will include all patients in the safety population, unless specifically stated otherwise.

In the rare situation where a patient is Lost to Follow-up at the first postbaseline visit, but some safety data exists (e.g., unscheduled laboratory assessments) after first dose of study drug, a listing of the data or a patient profile may be provided, when requested.

6.2.2. Definition of Baseline and Postbaseline Measures

The baseline value for efficacy and health outcomes variables measured at scheduled visits is defined as the last non-missing measurement on or prior to the date of first study drug administration (expected at Week 0, Visit 2).

The baseline value for the daily diary assessments (Itch NRS, ADSS, Skin Pain NRS, PGI-S-AD) is the mean of the non-missing assessments in the 7 days prior to the date of first study drug administration (expected at Week 0, Visit 2).

If there are less than 4 non-missing assessments in the baseline diary window, the interval lower bound can be extended up to 7 additional days, one day at a time, to obtain the most recent 4 non-missing values. If there are not at least 4 non-missing assessments in the baseline period, the baseline mean is missing.

Baseline for the safety analyses is defined as the last non-missing scheduled (planned) measurement on or prior to the date of first study drug administration for continuous measures by-visit analyses and all non-missing measurements on or prior to the date of first study drug administration for all other analyses.

Postbaseline measurements are collected after study drug administration through Week 16 (Visit 8) or early discontinuation visit. Efficacy data collected at scheduled visits (e.g., eCOA, ClinRO) will be used in all analyses unless it is missing. If an assessment is missing at a scheduled visit, an unscheduled post-baseline assessment can be used provided it falls within the

window interval as follows: a ±2 day window is used for Visit 3 (Week 1), Visit 4 (Week 2), and Visit 5 (Week 4); and a ±4 day window is used to Visit 6 (Week 8), Visit 7 (Week 12), and Visit 8 (Week 16). If there is more than 1 unscheduled visit within the defined visit window and no scheduled visit assessment is available, the unscheduled visit closest to the scheduled visit date will be used. If two unscheduled visits of equal distance are available, then the latter of the two will be used.

Postbaseline daily diary endpoints will be the mean of weekly visit windows (diary windows) anchored on day of first dose (Day 1) for Weeks 1 through 14 as follows: Week 1 (Days 1 through7), Week 2 (Days 8 through 14), Week 3 (Days 15 through 21), …, Week 14 (Days 92 through 98).

Week 16 Daily Diary Window Construction

The following sequential steps will be used to determine the Week 16 diary window. The general goal is to anchor on the scheduled Week 16 visit (or a proximal unscheduled visit) if such a visit exists or to use an interval based on days in study for cases where a scheduled Week 16 or a proximal surrogate does not exist.

Step 1: If the Week 16 scheduled visit exists, the Week 16 diary interval is the 7 days prior to the Week 16 date provided that window has at least 4 non-missing observations. If there are less than 4 non-missing observations, the diary window's lower bound will be extended 1 day at a time (up to day 99) to a maximum of 14 days prior to the Week 16 date until 4 non-missing observations are obtained. If, after extending this diary window's lower bound to 14 days, there are less than 4 non-missing observations then go to Step 2.

Step 2: If the Week 16 scheduled visit does not exist, the 7 days prior to the last visit (scheduled or unscheduled) occurring after Day 105, will constitute the Week 16 diary window provided that window contains at least 4 non-missing observations. If there are less than 4 non-missing observations, the diary window's lower bound will be extended 1 day at a time (up to Day 99) to a maximum of 14 days prior to the unscheduled visit date until 4 non-missing observations are obtained. If, after extending this diary window's lower bound to 14 days, there are less than 4 non-missing observations then go to Step 3.

Step 3: If neither a Week 16 scheduled visit is available nor an unscheduled visit to act as a surrogate for the Week 16 diary window, then the Week 16 window will be Day 106 to Day 112. If there are less than 4 non-missing observations, the dairy window's lower bound will be extended 1 day at a time to Day 99 until 4 non-missing observations are obtained.

If the steps above do not detect a window with at least 4 non-missing observations, then the Week 16 window is 7 days from either the Week 16 visit, the surrogate visit or Days 106 through 112 and the mean is missing and subject to imputation rules.

Week 15 Daily Diary Window Construction

The lower boundary of the Week 15 diary window is defined as Day 99. The upper bound of the Week 15 diary window is the minimum of either Day 105 or the lower bound of the Week 16

diary window -1. Consequently, Week 15 may be less than 4 days if the Week 16 scheduled visit is before Day 112. Moreover, as Week 15 diary window cannot exceed 7 days, there could be daily assessments between Weeks 15 and 16 diary windows that do not fall into a diary window. If after constructing the diary windows, there are fewer than 4 non-missing values the mean for Week 15 is missing and subject to imputation rules.

Handling of Duplicate Diary Records

If there is more than one diary record on a particular date, the first record on that particular date will be used in the analysis.

Note, as some analyses require use of the primary censoring rule, assessments collected on the day of rescue or afterwards will be excluded from the weekly visit interval calculation when implementing the rule for daily diary. If, after exclusion of these records, there are less than 4 non-missing assessments, the weekly interval which implements the primary censoring rule will be missing. The post-study follow-up weekly score for daily diaries will be calculated as the mean of the 7 days prior to the follow-up visit which occur after last dose of study treatment.

Postbaseline measures for the safety analyses are defined as the non-missing scheduled (planned) measurements after the date of first study drug administration for continuous measures by-visit analyses and all non-missing measurements after the date of first study drug administration for all other analyses.

6.2.3. Analysis Methods

The main analysis method of categorical efficacy variables and health outcomes variables will use a logistic regression analysis with region, baseline disease severity (IGA), baseline value and treatment group in the model, except for the analysis on IGA. For IGA, the logistic regression model will include region, baseline disease severity (IGA), and treatment group. Firth's correction will be used in order to accommodate (potential) sparse response rates. The p-value for the odds ratio from the logistic regression model will be used for statistical inference, unless Firth's correction still results in quasi-separation. In that case, Fisher's exact test will be used for statistical inference. The difference in percentages and 100(1-alpha)% CI of the difference in percentages using the Newcombe-Wilson method without continuity correction will be reported. The p-value from the Fisher's exact test will also be produced as a secondary analysis.

The main analysis method for all continuous efficacy and health outcomes variables will use mixed model repeated measures (MMRM) analysis. The MMRM model will use a restricted maximum likelihood (REML) estimation. The model will include treatment, region, baseline disease severity (IGA), visit, and treatment-by-visit-interaction as fixed categorical effects and baseline and baseline-by-visit-interaction as fixed continuous effects. For daily diary assessments, the model for analyses up to Week 16 will include all weekly assessments. An unstructured (co)variance structure will be used to model the between- and within-patient errors. If this analysis fails to converge, the heterogeneous autoregressive [ARH(1)], followed by the heterogeneous compound symmetry (CSH), followed by the heterogeneous Toeplitz (TOEPH), followed by autoregressive [AR(1)], followed by compound symmetry (CS) will be used. The

Kenward-Roger method will be used to estimate the degrees of freedom. Treatment least squares means (LSM) will be estimated within the framework of the MMRM using type 3 sums of squares. Differences in LSM between each dose of baricitinib and placebo (and associated p-values, standard errors and 95% CI) will be used for statistical inference. The LSM difference, standard error, p-value and 95% CI will be reported.

Treatment comparisons for continuous efficacy and health outcomes variables may also be made using analysis of covariance (ANCOVA) for primary and key secondary objectives. When an ANCOVA model is used, the model includes region, baseline disease severity, treatment group, and baseline value. Treatment LSM will be estimated within the framework of the ANCOVA using type 3 sums of squares. Reported differences in LSM and associated p-values, standard errors and 95% CI will be used for statistical inference. Treatment-by-region interaction will also be added to the model for sensitivity purposes and is discussed in Section 6.5.

Beginning on Day 14 a Cox proportional hazard (CPH) model of time (in days) to first observance of a 4-point itch reduction with effects for treatment, region, baseline mean itch, and disease severity will be used to test for treatment differences from placebo. For this analysis, daily itch scores will be compared to the baseline to determine if a 4-point itch reduction has been achieved in patients with a baseline itch of at least 4. The baseline for Itch NRS is defined in Section 6.2.2. Beginning on Day 14, the day on which the first time itch NRS is reduced by at least 4 will be modeled. If any significant difference between any baricitinib dose and placebo is observed then the same analyses will be run on Day 13. This process of evaluating at the next lowest day will proceed until no significant differences are observed. No adjustments for multiple tests and multiple comparisons will be used. This analysis uses the Primary Censoring Rule for patients who are rescued or permanently discontinue study drug (see Section 6.4). Missing daily itch data will be replaced using NRI rule which means missing data is replaced with a non-response which would entail replacing missing values with a time to event of >14 days, censored; >13 days, censored, etc., depending on the window being used. If the model assumptions for the CPH model do not hold, a log-rank test will be used.

Restricted mean survival time will be evaluated at Days 2, 3, 4 and 5 as an exploratory analysis. The model will include terms for baseline disease severity (IGA), baseline itch, region, and treatment group.

Fisher's exact test will be used to test for differences between each baricitinib dose and placebo in proportions of patients experiencing adverse events (AEs), discontinuation from study drug, and for other categorical safety data. Continuous vital signs, body weight, and other continuous safety variables, including laboratory variables will be analyzed by an ANCOVA with treatment group and baseline value in the model. The significance of within-treatment group changes from baseline will be evaluated by testing whether or not the treatment group LSM changes from baseline are different from zero; the standard error for the LSM change will also be displayed. Differences in LSM will be displayed, with the p-value associated with the LSM comparison to placebo and a 95% CI on the LSM difference also provided. In addition to the LSMs for each group, the within-group p-value for the change from baseline will be displayed.

Confusion table will be used to calculate the proportions of responder patients defined at week 16, and at any time between Weeks 8 and 16, for those with improvement on itch NRS or IGA at early week. The confusion table will generate the Positive Prediction Value (PPV) and Negative Prediction Value (NPV), which can access the performance of predicting late week responders using early week data. All missing values will be treated by NRI method.

6.2.4. Derived Data

- Age (year), derived using first dose date as the reference start date and July 1st of birth year and truncated to a whole-year (integer) age. Patients whose derived age is less than 18 will have the required minimum age of 18 at informed confirmed; reporting for age, age groups, and lab ranges, however will be based on their derived age.
- Age group (<65, ≥65 years old)
- Age group (<65, ≥65 to <75, ≥75 to <85, ≥85 years old)
- Body Mass Index (BMI) (kg/m^2) = Weight (kg)/((Height (cm)/100)^2)
- BMI category (<25 kg/m^2, ≥25 to <30 kg/m^2, ≥30 kg/m^2)
- The duration of AD from diagnosis (years) = [(Date of informed consent – Date of AD diagnosis)+1]/ 365.25.
    - If year of onset is missing, duration of AD will be set as missing. Otherwise, unknown month will be taken as January, and unknown day will be taken as 01. The duration of AD will be rounded to 1 decimal place.
- Duration of AD (years) category (0 to <2 years, 2 to < 5 years, 5 to <10 years, 10 to <20 years, ≥20 years)
- Diagnosis age (years), derived using diagnosis date as the reference start date and July 1 of birth year and truncated to a whole-integer age.
- Diagnosis age group (<18, ≥18 to <50, ≥50 years old)
- Change from baseline = postbaseline measurement at Visit x – baseline measurement.
    - If a baseline value is missing, it will not be imputed and the change from baseline will not be calculated.
- Percent change from baseline at Visit x:
  ((Post-baseline measurement at Visit x - Baseline measurement)/Baseline measurement)*100.
    - If a baseline value is missing, it will not be imputed and percent change from baseline will not be calculated.
- Weight (kg) = weight (lbs) * 0.454.
- Weight category (<60 kg, ≥60 to <100 kg, ≥100 kg)
- Height (cm) = height (in) * 2.54.
- Cyclosporine inadequate efficacy (yes, no)
    - Set **yes** if the reason for discontinuation is inadequate response.
- Cyclosporine intolerance (yes, no)
    - Set **yes** if the reasons for discontinuation are: intolerance to medication or contraindication (Physician indicated cyclosporine was used and a contraindication was noted).
- Cyclosporine contraindication [ineligible] (yes, no)

    - Set to **yes** if cyclosporine never used because of a contraindication
- Cyclosporine inadvisable (yes, no)
  - Set to **yes** if the following reasons were selected for either not using the medication or discontinuing the medication:
    - Reason for not using medication: Physician decision, concern about side effects, unfavorable benefit risk, contraindication.
    - Reasons for discontinuation: inadequate response, intolerance to medication, or contraindication.
- TCNI inadequate efficacy (yes, no)
    - Set **yes** if the reason for discontinuation is inadequate response.
- TCNI intolerance (yes, no)
    - Set **yes** if the reasons for discontinuation are: intolerance to medication or contraindication (Physician indicated TCNI was used and a contraindication was noted).
- TCNI contraindication / [ineligible](yes, no)
    - Set to **yes** if TCNI never used because of a contraindication
- TCNI inadvisable (yes, no)
  - Set to **yes** if the following reasons were selected for either not using the medication or discontinuing the medication:
    - Reason for not using medication: Physician decision, concern about side effects, unfavorable benefit risk, contraindication.
    - Reasons for discontinuation: inadequate response, intolerance to medication, or contraindication.

6.3. Covariate Adjustment

The randomization to treatment groups at Week 0 (Visit 2) is stratified by disease severity (IGA) and geographic region as described in Section 5.1. Unless otherwise specified, the statistical analysis models will adjust for these stratification variables. The covariates used in the logistic model for categorical data will include the parameter value at baseline. The covariates used in the ANCOVA model for continuous data will include the parameter value at baseline. Inclusion of baseline in the model ensures treatment LSM are estimated at the same baseline value. When an MMRM analysis is performed, baseline value and baseline-by-visit interactions will be included as covariates.

6.4. Handling of Dropouts or Missing Data

Intercurrent events (ICH E9 R1) are events which occur after the treatment initiation and make it impossible to measure a variable or influence how it would be interpreted.

Depending on the estimand being addressed, different methods will be used to handle missing data as a result of intercurrent events. Intercurrent events can occur through the following:

- application of one of the censoring rules (including after permanent study drug discontinuation or after rescue therapy)
- discontinuation

- missing an intermediate visit prior to discontinuation or rescue
- lost to follow-up.

Non-censor intercurrent events are events that are not due to the application of any censoring rule, i.e., the last 3 items in the list above.

Note that as efficacy and health outcome data can accrue after a patient permanently discontinues study drug or begins rescue therapy, specific general censoring rules to the data will be applied to all efficacy and health outcome observations subsequent to these events depending on the estimand being addressed. These specific censoring rules are described below.

The *primary censoring rule* will censor efficacy and health outcome data after permanent study drug discontinuation or after rescue therapy. This censoring rule will be applied to all continuous and categorical efficacy and health outcome endpoints. This censoring rule is equivalent to using all the data up to rescue.

A *secondary censoring rule* will only censor efficacy and health outcome data after permanent study drug discontinuation. This sensitivity analysis will include all observed values up to study drug discontinuation. The secondary censoring rule will be applied to primary and key secondary efficacy and health outcome endpoints as sensitivity analyses.

Table JAIY.6.1 describes the planned imputation methods for efficacy and health outcome endpoints with associated censoring rules. Sections 6.4.1 through 6.4.5 summarize the methodology of each imputation rule.

Table JAIY.6.1. Imputation Techniques for Various Variables

Efficacy and Health Outcome Endpoints	Imputation Method
IGA(0,1), EASI75, 4-point Itch NRS improvement, EASI90, SCORAD75	NRI[ab], pMI[a], Tipping point[a]
EASI percent change, ADSS Item 2 change, Skin Pain NRS change	MMRM[ab], mLOCF[a], pMI[a]
All remaining categorical measures	NRI[a]
All remaining continuous efficacy and health outcome measures	MMRM[a], mLOCF[a]

Abbreviations: ADSS = Atopic Dermatitis Sleep Scale; EASI = Eczema Area and Severity Index score; IGA = Investigator's Global Assessment for AD; mLOCF = modified last observation carried forward; MMRM = mixed model repeated measures; NRI = nonresponder imputation, NRS = Numeric Rating Scale; pMI = placebo multiple imputation; SCORAD = SCORing Atopic Dermatitis.

a Analyses utilizing the primary censoring rule.

b Analyses utilizing the secondary censoring rule.

6.4.1. Nonresponder Imputation

A nonresponder imputation (NRI) method imputes missing values as non-responses and can be justified based on the composite strategy for handling intercurrent events (ICH E9 R1). This imputation procedure assumes the effects of treatments disappear after the occurrence of an intercurrent event defined by the associated censoring rule.

All categorical endpoints will utilize the NRI method after applying the primary censoring rule to patients who permanently discontinued study drug or were rescued (described in Section 6.4). Additionally, all primary and key secondary categorical endpoints will utilize NRI after applying the secondary censoring rule as sensitivity analyses. For analyses which utilize either of the censoring methods, randomized patients without at least 1 post-baseline observation will be defined as nonresponders for all visits.

6.4.2. Mixed Model for Repeated Measures

Mixed Model for Repeated Measures analyses will be performed on continuous endpoints to mitigate the impact of missing data. This approach assumes missing observations are missing-at-random (missingness is related to observed data) and borrows information from patients in the same treatment arm taking into account both the missingness of data through the correlation of the repeated measurements.

Essentially MMRM estimates the treatment effects had all patients remained on their initial treatment throughout the study. For this reason, the MMRM implies a different estimand (hypothetical strategy [ICH E9 R1]) than the one used for NRI on categorical outcomes.

All continuous endpoints will utilize MMRM after applying the primary censoring rule. As sensitivity analyses, all secondary continuous endpoints will also utilize MMRM after applying the secondary censoring rule (Table JAIY.6.1).

6.4.3. Modified Last Observation Carried Forward

For continuous measure, a modified last observation carried forward (mLOCF) imputation technique replaces missing data with the most recent non-missing post-baseline assessment. The specific modification to the LOCF is data after an intercurrent event will not be carried forward thus the mLOCF is applied after the specified censoring rule is implemented. The mLOCF assumes the effect of treatment remain the same after the event that caused missing data as it was just prior to the missing data event. Analyses using mLOCF require a nonmissing baseline and at least 1 postbaseline measure otherwise the data is missing for analyses purposes. Analyses using mLOCF help ensure the number of randomized patients who were assessed post-baseline is maximized and is reasonable for this data as data directly prior to an intercurrent event (such as initiation of rescue therapy or drop out) is likely a non-efficacious response.

All continuous efficacy and health outcomes endpoints will use with mLOCF imputation methodology with an ANCOVA as sensitivity analyses to the MMRM analyses.

6.4.4. Placebo Multiple Imputation

The Placebo Multiple Imputation (pMI) methodology will be used as a sensitivity analysis for the analysis of the primary efficacy endpoint (IGA 0 or 1 at Week 16) as well as the key secondary endpoints at Week 16. In these sensitivity analyses the primary censoring rule will be applied.

The pMI assumes that the statistical behavior of drug- and placebo- treated patients after the occurrence of intercurrent events will be the same as if patients were treated with placebo. Thus,

in the effectiveness context, pMI assumes no pharmacological benefit of the drug after the occurrence of intercurrent events but is a more conservative approach than mLOCF because it accounts for uncertainty of imputation, and therefore does not underestimate standard errors, and it limits bias. In the efficacy context pMI is a specific form of a missing not at random analysis and expected to yield a conservative estimate of efficacy.

In the pMI analysis, multiple imputations are used to replace missing outcomes for drug- and placebo-treated patients who have an intercurrent event using multiple draws from the posterior predictive distribution estimated from the placebo arm. The binary outcomes will then be derived from the imputed data.

Data are processed sequentially by repeatedly calling SAS® PROC MI to impute missing outcomes at visits t=1,.., T.

1. *Initialization:* Set t=0 (baseline visit)
2. *Iteration:* Set t=t+1. Create a data set combining records from drug- and placebo-treated patients with columns for covariates **X** and outcomes at visits 1,..,t with outcomes for all drug-treated patients set to missing at visit t and set to observed or imputed values at visits 1,..,t-1.
3. *Imputation:* Run Bayesian regression in SAS® PROC MI on this data to impute missing values for visit t using previous outcomes for visits 1 to t-1 and baseline covariates. Note that only placebo data will be used to estimate the imputation model since no outcome is available for drug-treated patients at visit t.
4. Replace imputed data for all drug-treated patients at visit t with their observed values, whenever available up to permanent study drug discontinuation and/or rescue (if censoring on rescue). If t < T then go to Step 2, otherwise proceed to Step 5.
5. Repeat steps 1-4, m times with different seed values to create m imputed complete data sets.

Analysis: For continuous endpoints, fit its treatment response model (MMRM) for each completed data set. For the primary and secondary key efficacy endpoints [IGA (0,1), EASI75, EASI90, SCORAD75, and 4-point improvement from baseline in Itch NRS], the binary outcomes will be derived from the imputed data for each patient before fitting the logistic regression model.

The number of imputed data sets will be m=100 and a 6-digit seed value will be pre-specified for each analysis. Within the program, the seed will be used to generate the m seeds needed for imputation. The initial seed values are given in Table JAIY.6.2.

Table JAIY.6.2. Seed Values for Multiple Imputation

Analysis	Seed value
Proportion of patients achieving IGA of 0 or 1 with a ≥2-point improvement from baseline at Week 16 using the primary censoring rule	123450
Percent change from baseline in EASI score at 16 weeks using the primary censoring rule. EASI75 and EASI90 will leverage imputation from EASI and therefore do not need a new seed number.	123451
Proportion of patients achieving SCORAD75 at 16 week using the primary censoring rule, with data up to rescue	123452
Proportions of patients achieving a 4-point improvement from baseline in Itch NRS at Week 16 using the primary censoring rule	123453
Mean change from baseline in Skin Pain NRS at Week 16 using the primary censoring rule	123454
Mean change from baseline in the score of Item 2 of the ADSS at Week 16 using the primary censoring rule	123455

Abbreviations: ADSS = Atopic Dermatitis Sleep Scale; EASI = Eczema Area and Severity Index score; IGA = Investigator's Global Assessment for AD; NRS = Numeric Rating Scale; SCORAD = SCORing Atopic Dermatitis.

The final inference on treatment difference is conducted from the multiple datasets using Rubin's combining rules, as implemented in SAS® PROC MIANALYZE.

6.4.5. Tipping Point Analyses

To investigate the missing data mechanism, sensitivity analyses using multiple imputation (MI) under the missing not at random assumption will be provided for the following primary and key secondary objectives as given in Table JAIY.6.3.

All patients in the ITT population will be included. Data after the occurrence of intercurrent events (after application of the primary censoring rule) will be set to missing.

Within each analysis, a most extreme case will be considered, in which all missing data for patients randomized to baricitinib 2 mg or 4 mg will be imputed using the worst possible result and all missing data for patients randomized to placebo will be imputed with the best possible result. Treatment differences will be analyzed using logistic regression or ANCOVA (Section 6.1) as appropriate.

For continuous variables, the following process will be used to determine the tipping point:

1. To handle intermittent missing visit data, a Markov chain Monte Carlo method (SAS® Proc MI with MCMC option) will be used to create a monotone missing pattern.
2. A set of Bayesian regressions (using SAS® Proc MI with MONOTONE option) will be used for the imputation of monotone dropouts. Starting from the first visit with at least 1 missing value, the regression models will be fit sequentially with treatment as a fixed effect and values from the previous visits as covariates.

3. A delta score is added to all imputed scores at the primary time point for patients in the baricitinib treatment groups, thus worsening the imputed value. The delta score is capped for patients based on the range of the outcome measure being analyzed.
4. Treatment differences between baricitinib and placebo are analyzed for each imputed dataset using ANCOVA (Section 6.1). Results across the imputed datasets are aggregated using SAS® Proc MIANALYZE in order to compute a p-value for the treatment comparisons for the given delta value.
5. Steps 3 and 4 are repeated, and the delta value added to the imputed baricitinib scores is gradually increased. The tipping point is identified as the delta value which leads to a loss of statistical significance (aggregated p-value >0.05) when evaluating baricitinib relative to the placebo group.

As a reference, for each delta value used in Steps 3 through 5, a fixed selection of delta values (ranging from slightly negative to slightly positive) will be added to imputed values in the placebo group, and Step 4 will be performed for the combination. This will result in a 2-d table, with the columns representing the delta values added to the imputed placebo responses, and the rows representing the delta values added to the imputed baricitinib responses. Separate 2-d tables will compare each baricitinib dose group to placebo.

A similar process will be used for the categorical variables:

1. Missing responses in the baricitinib groups will be imputed with a range of low response probabilities, including probabilities of 0, 0.1, and 0.2.
2. For missing responses in the placebo group, a range of responses probabilities (for example, probability = 0, 0.2 … 1) will be used to impute the missing values. Multiple imputed datasets will be generated for each response probability.
3. Treatment differences between baricitinib and placebo are analyzed for each imputed dataset using logistic regression (Section 6.1). Results across the imputed datasets are aggregated using SAS® Proc MIANALYZE in order to compute a p-value for the treatment comparisons for the given response probability. If the probability values do not allow for any variation between the multiple imputed datasets (for example, all missing responses in the placebo and baricitinib groups are imputed as responders and nonresponders, respectively), then the p-value from the single imputed dataset will be used.

The tipping point is identified as the response probability value within the placebo group that leads to a loss of statistical significance when evaluating baricitinib relative to placebo.

For tipping point analyses the number of imputed data sets will be m=100 and the seed values to start the pseudorandom number generator of SAS Proc MI (same values for MCMC option and for MONOTONE option) are given in Table JAIY.6.3.

Table JAIY.6.3. Seed Values for Imputation

Analysis	Seed value
Proportion of patients achieving IGA (0,1) with ≥2-point improvement at Week 16; primary censoring rule	123470
Proportion of patients achieving EASI75 at Week 16; primary censoring rule Proportion of patients achieving EASI90 at Week 16; primary censoring rule	123471
Proportions of patients achieving a 4-point improvement from baseline in Itch NRS at Week 16, primary censoring rule	123472
Proportion of patients achieving SCORAD75 at Week 16; primary censoring rule	123473

6.5. Multicenter Studies

This study will be conducted by multiple investigators at multiple sites internationally. The countries will be categorized into geographic regions, as described in Section 5.2.

For the analysis of the primary endpoint, treatment-by-region interaction will be added to the logistic regression model as a sensitivity analysis and results from this model will be compared to the primary model (without the interaction effect). If the treatment-by-region interaction is significant at a 2-sided α level of 0.1, the nature of this interaction will be inspected as to whether it is quantitative (i.e., the treatment effect is consistent in direction across all regions but not in size of treatment effect) or qualitative (the treatment is beneficial in some but not all regions). If the treatment-by-region interaction effect is found to be quantitative, results from the primary model will be presented. If the treatment-by-region interaction effect is found to be qualitative, further inspection will be used to identify in which regions baricitinib is found to be more beneficial.

6.6. Multiple Comparisons/Multiplicity

The primary and key secondary endpoints will be adjusted for multiplicity in order to control the overall family-wise Type I error rate at a 2-sided alpha level of 0.05.

The following is a list of primary and key secondary endpoints to be tested.

Primary Null Hypotheses:

- Null Hypotheses[IGA0-1]: Proportion of baricitinib 4-mg patients achieving IGA of 0 or 1 with a ≥2-point improvement from baseline at Week 16 is equal to the proportion of placebo patients achieving IGA of 0 or 1 with a ≥2-point improvement from baseline at Week 16
- Null Hypotheses[IGA0-1]: Proportion of baricitinib 2-mg patients achieving IGA of 0 or 1 with a ≥2-point improvement from baseline at Week 16 is equal to the proportion of placebo patients achieving IGA of 0 or 1 with a ≥2-point improvement from baseline at Week 16

Key Secondary Null Hypotheses:

- Null Hypotheses[EASI75]: Proportion of baricitinib 4-mg patients achieving EASI75 is equal to the proportion of placebo patients achieving EASI75 at Week 16
- Null Hypotheses[ITCH W16]: Proportion of baricitinib 4-mg patients achieving a 4-point improvement in Itch NRS is equal to the proportion of placebo patients achieving a 4-point improvement in Itch NRS at Week 16 among patients with baseline Itch NRS score ≥4
- Null Hypotheses[EASI PCFB]: Percent change from baseline in EASI score for baricitinib 4-mg patients is equal to the percent change from baseline in EASI score for placebo patients at Week 16
- Null Hypotheses[ITCH W4]: Proportion of baricitinib 4-mg patients achieving a 4-point improvement in Itch NRS is equal to the proportion of placebo patients achieving a 4-point improvement in Itch NRS at Week 4 among patients with baseline Itch NRS score ≥4
- Null Hypotheses[SCORAD75]: Proportion of baricitinib 4-mg patients achieving SCORAD75 is equal to the proportion of placebo patients achieving SCORAD75 at Week 16
- Null Hypotheses[EASI 90]: Proportion of baricitinib 4-mg patients achieving EASI90 is equal to the proportion of placebo patients achieving EASI90 at Week 16
- Null Hypotheses[PAIN NRS]: Mean change from baseline in Skin Pain NRS for baricitinib 4-mg patients is equal to the mean change from baseline in Skin Pain NRS for placebo patients at Week 16
- Null Hypotheses[ADSS2 W16]: Mean change from baseline in the score of Item 2 of the ADSS for baricitinib 4-mg patients equal to the mean change from baseline in the score of Item 2 of the ADSS for placebo patients at Week 16
- Null Hypotheses[ITCH W2]: Proportion of baricitinib 4-mg patients achieving a 4-point improvement in Itch NRS is equal to the proportion of placebo patients achieving a 4-point improvement in Itch NRS at Week 2 among patients with baseline Itch NRS score ≥4
- Null Hypotheses[ADSS2 W1]: Mean change from baseline in the score of Item 2 of the ADSS for baricitinib 4-mg patients is equal to the mean change from baseline in the score of Item 2 of the ADSS for placebo patients at Week 1
- Null Hypotheses[ITCH W1]: Proportion of baricitinib 4-mg patients achieving a 4-point improvement in Itch NRS is equal to the proportion of placebo patients achieving a 4-point improvement in Itch NRS at Week 1 among patients with baseline Itch NRS score ≥4
- Null Hypotheses[ITCH D2]: Proportion of baricitinib 4-mg patients achieving a 4-point improvement in Itch NRS is equal to the proportion of placebo patients achieving a 4-point improvement in Itch NRS at Day 2 among patients with baseline Itch NRS score ≥4
- Null Hypotheses[EASI75]: Proportion of baricitinib 2-mg patients achieving EASI75 is equal to the proportion of placebo patients achieving EASI75 at Week 16
- Null Hypotheses[ITCH W16]: Proportion of baricitinib 2-mg patients achieving a 4-point improvement in Itch NRS is equal to the proportion of placebo patients achieving a 4-point improvement in Itch NRS at Week 16 among patients with baseline Itch NRS score ≥4

- Null Hypotheses[EASI PCFB]: Percent change from baseline in EASI score for baricitinib 2-mg patients is equal to the percent change from baseline in EASI score for placebo patients at Week 16
- Null Hypotheses[ITCH W4]: Proportion of baricitinib 2-mg patients achieving a 4-point improvement in Itch NRS is equal to the proportion of placebo patients achieving a 4-point improvement in Itch NRS at Week 4 among patients with baseline Itch NRS score ≥4
- Null Hypotheses[SCORAD75]: Proportion of baricitinib 2-mg patients achieving SCORAD75 is equal to the proportion of placebo patients achieving SCORAD75 at Week 16
- Null Hypotheses[EASI90]: Proportion of baricitinib 2-mg patients achieving EASI90 is equal to the proportion of placebo patients achieving EASI90 at Week 16
- Null Hypotheses[PAIN NRS]: Mean change from baseline in Skin Pain NRS for baricitinib 2-mg patients is equal to the mean change from baseline in Skin Pain NRS for placebo patients at Week 16
- Null Hypotheses[ADSS2 W16]: Mean change from baseline in the score of Item 2 of the ADSS for baricitinib 2-mg patients is equal to the mean change from baseline in the score of Item 2 of the ADSS for placebo patients at Week 16
- Null Hypotheses [ITCH W2]: Proportion of baricitinib 2-mg patients achieving a 4-point improvement in Itch NRS is equal to the proportion of placebo patients achieving a 4-point improvement in Itch NRS at Week 2 among patients with baseline Itch NRS score ≥4
- Null Hypotheses[ADSS2 W1]: Mean change from baseline in the score of Item 2 of the ADSS for baricitinib 2-mg patients is equal to the mean change from baseline in the score of Item 2 of the ADSS for placebo patients at Week 1
- Null Hypotheses[ITCH W1]: Proportion of baricitinib 2-mg patients achieving a 4-point improvement in Itch NRS is equal to the proportion of placebo patients achieving a 4-point improvement in Itch NRS at Week 1 among patients with baseline Itch NRS score ≥4
- Null Hypotheses[ITCH D2]: Proportion of baricitinib 2-mg patients achieving a 4-point improvement in Itch NRS is equal to the proportion of placebo patients achieving a 4-point improvement in Itch NRS at Day 2 among patients with baseline Itch NRS score ≥4

A multiple testing strategy for the primary and key secondary endpoints is implemented through a graphical testing scheme depicted by Figure JAIY.6.1.

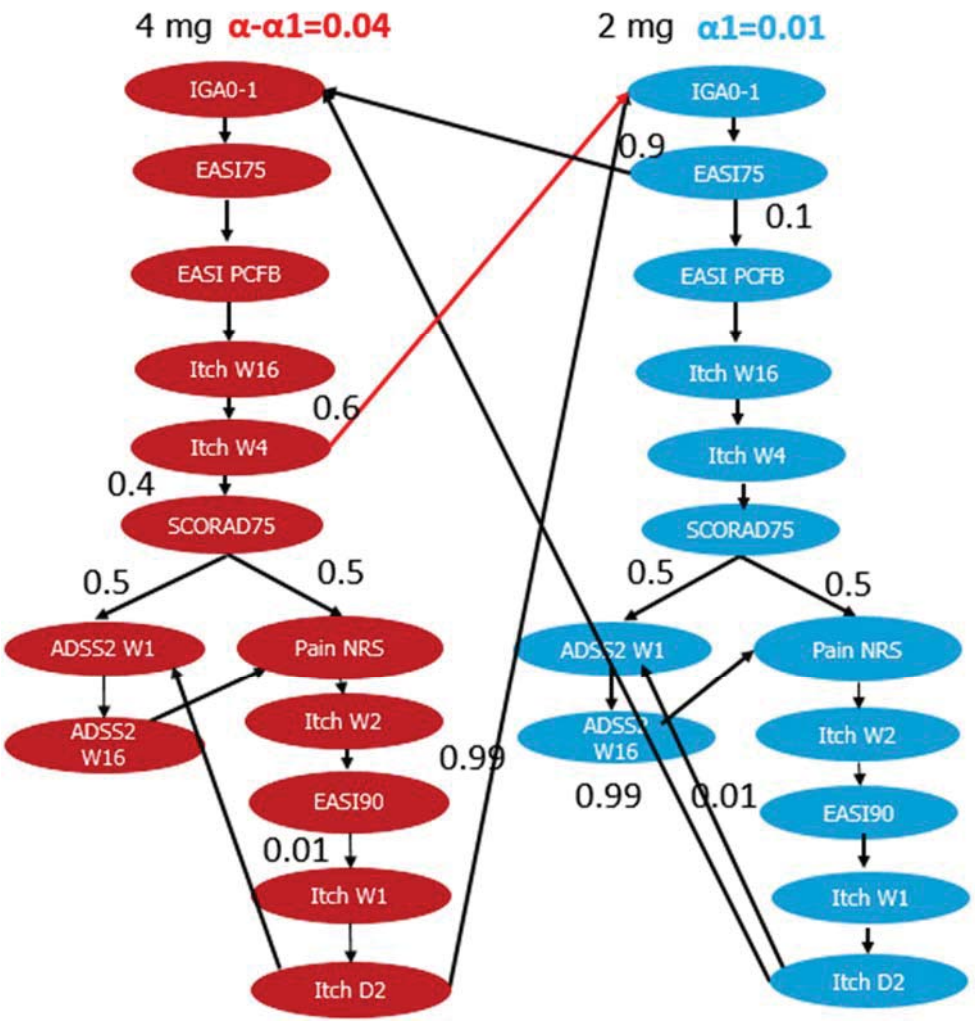


Figure JAIY.6.1. Illustration of graphical multiple testing procedure with initial α allocation and weights.

There will be no adjustment for multiple comparisons for any other analyses.

6.7. Patient Disposition

An overview of patient populations will be summarized by treatment group. Frequency counts and percentages of patients excluded prior to randomization by primary reason for exclusion will be provided for patients who failed to meet study entry requirements during screening.

Patient disposition through Week 16 will be summarized using the ITT population. Frequency counts and percentages of patients who complete the study treatment visits or discontinue early from the study along with whether they completed follow-up, did not complete follow-up or enrolled into the extension will be summarized separately by treatment group for patients who are not rescued and for patients who are rescued, along with their reason for study discontinuation. Frequency counts and percentages of patients who complete the treatment or discontinue treatment early will also be summarized separately by treatment group for patients

who are not rescued and for patients who are rescued, along with their reason for treatment discontinuation.

A listing of patient disposition will be provided for all randomized patients, with the extent of their participation in the study and the reason for discontinuation. A listing of all randomized patients with their treatment assignment will also be provided.

6.8. Patient Characteristics

Patient characteristics including demographics and baseline characteristics will be summarized descriptively by treatment group for the ITT population. Historical illnesses and pre-existing conditions will be summarized descriptively by treatment group for the ITT population. No formal statistical comparisons will be made among treatment groups unless otherwise stated.

6.8.1. Demographics

Patient demographics will be summarized as described above. The following demographic information will be included:

- Age
- Age group (<65 vs. ≥65)
- Age group (<65, ≥65 to <75, ≥75 to <85, ≥85)
- Gender (male, female)
- Race (American Indian or Alaska Native, Asian, Black or African American, Native Hawaiian or Other Pacific Islander, White, Multiple)
- Region (as defined in Table JAIY.5.1)
- Country
- Weight (kg)
- Weight category (<60 kg, ≥60 to <100 kg, ≥100 kg)
- Height (cm)
- BMI (kg/m^2)
- BMI category (<25 kg/m^2, ≥25 to <30 kg/m^2, ≥30 kg/m^2)

A listing of patient demographics will also be provided for the ITT population.

6.8.2. Baseline Disease Characteristics

The following baseline disease information will be categorized and presented for baseline AD clinical characteristics, baseline health outcome measures, and other baseline demographic and disease characteristics as described above:

- Duration since AD diagnosis (years)
- Duration since AD diagnosis category (0 to <2 years, 2 to <5 years, 5 to <10 years, 10 to <20 years, ≥20 years)
- Age at Diagnosis (years)
- Age Group at Diagnosis (<18 years, ≥18 to <50 years, ≥50 years)
- Habits (Alcohol: Never, Current, Former; Tobacco: Never, Current, Former)

- Skin Infections treated with a pharmacological agent within past year (yes, no, unknown; number if yes)
- Atopic Dermatitis Flares within past year (yes, no, unknown; number if yes)
- Validated Investigator's Global Assessment for AD (IGA) score
- Eczema Area and Severity Index (EASI) score
- SCORing Atopic Dermatitis (SCORAD)
- Body Surface Area (BSA) affected by AD
- Hospital Anxiety Depression Scale (HADS) subscales
- Patient-Oriented Eczema Measure (POEM)
- Itch Numerical Rating Scale (NRS)
- Atopic Dermatitis Sleep Scale (ADSS) Item 2
- Dermatology Life Quality Index (DLQI)
- Skin Pain NRS
- Patient Global Impression of Severity (PGI-S-AD)
- Prior therapy (topical therapy only; systemic therapy)
- Prior use of Cyclosporine (yes, no)
- Cyclosporine inadequate response (yes, no)
- Cyclosporine intolerance (yes, no)
- Cyclosporine contraindication [ineligible] (yes, no)
- Cyclosporine inadvisable (yes no)
- Prior use of TCNI (yes, no)
- TCNI inadequate response (yes, no)
- TCNI intolerance (yes, no)
- TCNI contraindication [ineligible] (yes, no)
- TCNI inadvisable (yes, no)
- Vaccine (yes, no)
- Baseline renal function status: impaired (eGFR <60 mL/min/1.73 m^2) or not impaired (eGFR ≥60 mL/min/1.73 m^2)
- Immunoglobulin E (IgE): intrinsic(<200 kU/I) or extrinsic (≥200 kU/I)

6.8.3. Historical Illness and Pre-existing Conditions

Historical illnesses are defined as those conditions recorded in the Pre-existing Conditions and Medical History electronic case report form (eCRF) or from the Prespecified Medical History: Comorbidities eCRF with an end date prior to the informed consent date. The number and percentage of patients with selected historical diagnoses will be summarized by treatment group using the ITT population. Historical diagnoses will be categorized using the Medical Dictionary for Regulatory Activities (MedDRA®, most current available version) algorithmic standardized MedDRA queries (SMQs) or similar pre-defined lists of preferred terms (PTs) of interest.

Preexisting conditions are defined as those conditions with a start date prior to the first dose of the study drug and stop dates that are at or after the informed consent date or have no stop date (i.e., are ongoing). For events occurring on the day of the first dose of study treatment, the date and time of the onset of the event will both be used to determine if the event was pre-existing. Conditions with a partial or missing start date (or time if needed) will be assumed to be 'not pre-

existing' unless there is evidence, through comparison of partial dates, to suggest otherwise. Pre-existing conditions will be categorized using the MedDRA SMQs or similar pre-defined lists of PTs of interest. Frequency counts and percentages of patients with selected pre-existing conditions will be summarized by treatment group using the ITT population.

6.9. Treatment Compliance

Patient compliance with study medication will be assessed from Week 0 (Visit 2) to Week 16 (Visit 8) or Early Termination using the ITT population.

All patients are expected to take 2 tablets daily from a package as described in the protocol. Each bottle contains 36 tablets. A patient is considered noncompliant if he or she misses >20% of the prescribed doses during the study, unless the patient's study drug is withheld by the investigator. For patients who had their treatment temporarily interrupted by the investigator, the period of time that dose was withheld will be taken into account in the compliance calculation.

Compliance in the period of interest up to Visit *x* will be calculated as follows:

$$\text{Compliance} = \frac{\text{total number of tablets dispensed} - \text{total number of tablets returned}}{\text{expected number of total tablets}}$$

where

- Total number of tablets dispensed: sum of tablets dispensed in the period of interest prior to Visit *x*;
- Total number of tablets returned: sum of the tablets returned in the period of interest prior to and including Visit *x*;
- Expected number of tablets: number of days in the period of interest*number of tablets taken per day = [(date of last dose – date of first dose + 1) – number of days of temporary drug interruption]*number of tablets taken per day

Patients who are significantly noncompliant (compliance <80%) through Week 16 will be excluded from the PPS population.

Descriptive statistics for percent compliance and non-compliance rate will be summarized for the ITT population by treatment group for Week 0 through Week 16. Sub-intervals of interest, such as compliance between visits, may also be presented. The number of expected doses, tablets dispensed, tablets returned, and percent compliance will be listed by patient for Week 0 through Week 16.

6.10. Rescue Therapy

Rescue therapy with additional topical and systemic therapies is available starting after 2 weeks of treatment (Visit 4), for patients who are experiencing worsening and unacceptable symptoms of AD despite treatment with IP and moderate-potency TCS. Patients whose lesions persist or worsen despite the use of emollients and background TCS therapy (e.g., triamcinolone 0.1% cream and/or hydrocortisone 2.5% ointment) and/or patients who require prolonged applications of triamcinolone 0.1% cream (moderate-potency TCS) on large surfaces may be considered for rescue to high- or ultra-high-potency TCS. If topical rescue therapy as described above fails to sufficiently control AD symptoms, then oral systemic medications may be used as rescue (e.g.,

corticosteroids, cyclosporine, methotrexate); however, investigational product will be required to be permanently discontinued for the remainder of the 16-week study duration. If these medications are needed for other medical conditions (e.g., asthma flare), they will still be treated as rescue medications.

The initial rescue therapy will be the first non-missing record before the last dose date from the CRF page *Concomitant Therapy: Rescue Therapy*.

A summary of the initial rescue therapy and the reason for requiring initial rescue will be produced, as well as a summary of the proportion of patients initially rescued at each study visit. A summary of all rescue medications will be provided.

6.11. Previous and Concomitant Therapy

Summaries of previous and concomitant medications will be based on the ITT population.

At screening, previous and current AD treatments are recorded for each patient. Concomitant therapy for the treatment period is defined as therapy that starts before or during the treatment period and ends during the treatment period or is ongoing (has no end date or ends after the treatment period). Should there be insufficient data to make this comparison (for example, the concomitant therapy stop year is the same as the treatment start year, but the concomitant therapy stop month and day are missing), the medication will be considered as concomitant for the treatment period.

Summaries of previous medications will be as follows:

- Previous AD therapies

Summaries of concomitant medications, with sponsor and non-sponsor provided background TCS included, will be as follows:

- General Concomitant medications excluding rescue medicine

6.11.1. Background TCS

Background TCS therapy with moderate-potency and/or low-potency TCS (e.g., triamcinolone 0.1% cream and/or hydrocortisone 2.5% ointment) are to be used on active lesions, as described in Section 7.7.3 in the protocol.

The dispensed weight of sponsor-provided TCS tubes for the two different potencies (low and moderate) varies between countries due to different supply regions. Average weights of full tubes were used to determine the dispensed weights for each region. Returned tubes were weighed with cap (without the carton) to determine the amount of TCS in grams (g) used at each visit.

For low potency TCS, the dispensed tube weight with cap (without the carton) in Japan is 13.5g. For countries supplied by European distributors (Austria, Germany, Italy, Poland, and Spain), the dispensed weight of low potency TCS is 21g. The remaining countries, supplied by US distributors (Argentina, Australia, Korea, and Taiwan), the weight of low potency TCS is 40g.

For moderate potency TCS, the dispensed tube weight with cap (without the carton) in Japan is 13.5g. For countries supplied by European distributors, the dispensed weight of moderate potency TCS is 38g. The remaining countries, supplied by US distributors, the weight of moderate potency TCS is 40g.The total amount of background TCS, provided by sponsor, will be summarized in grams by potency (low and moderate) and both potencies, between visits (Week 0 through Week 1, Week 1 through Week 2, Week 2 through Week 4, Week 4 through Week 8, Week 8 through 12, Week 12 through Week 16), and throughout the entire 16-week treatment period. If a returned tube is not weighed or not returned, then the tube can be classified as partially used, fully used, unused, or unknown. Partially used rescue medication tubes will be defined as 50% used whereas fully used and unused tubes will be defined as 100% and 0% used respectively. When drug accountability is not performed for a particular tube of rescue medication or an answer of 'unknown' is given for a tube which is not returned, that particular tube will not be included in the analysis. The main analysis on the total amount of background TCS throughout the entire 16-week treatment period will apply censoring rule #1. After patients who get rescued or discontinue IP, whichever is earlier, it is assumed that they would use the same amount of TCS as they did before. Analysis will be done via analysis of variance (ANOVA), with geographic region, baseline disease severity and treatment as factors in the model. The secondary analysis will apply censoring rule #2 with the same assumptions as described above.

Whether any background TCS is used or not used for each patient is also collected on the diary device in each day starting from the first dose date throughout the study.

The total number of days that the patients did not use background TCS will be summarized by both potencies throughout the entire 16-week treatment period. The main analysis applies censoring rule #1. After patients who are rescued or discontinue IP, it is assumed that background TCS would be applied each day. In case of missing values in the daily diary, it will be assumed that background TCS has been used. Analysis will be done via ANOVA, with geographic region, baseline disease severity and treatment as factors in the model. A secondary analysis will apply censoring rule #2, with the same assumptions for missing values as described above.

6.12. Efficacy Analyses

The general methods used to summarize efficacy data, including the definition of baseline value for assessments are described in Section 6.2. The censoring rules applied to data as well as imputation methods are described in Section 6.4.

Table JAIY.6.4 provides the descriptions and derivations of the primary, secondary, and exploratory efficacy outcomes.

Table JAIY.6.5 provides the detailed analyses including analysis type, method and imputation, population, time point, and comparisons for efficacy analyses.

Table JAIY.6.4. Description and Derivation of Primary, Secondary and Exploratory Efficacy Outcomes

Measure	Description	Variable	Derivation / Comment	Imputation Approach if Missing Components
Validated Investigator's Global Assessment for AD (IGA)	The validated Investigator's global assessment of the patient's overall severity of their AD, based on a static, numeric 5-point scale from 0 (clear) to 4 (severe). The score is based on an overall assessment of the degree of erythema, papulation/induration, oozing/crusting, and lichenification.	IGA score	Single item. Range: 0 to 4 0 represents "clear" 4 represents "severe"	Single item, missing if missing.
		Change from baseline in IGA score	Change from baseline: observed IGA score – baseline IGA score	Missing if baseline or observed value is missing.
		▪ IGA [0,1] with ≥2-point improvement ▪ IGA [0]	▪ Observed score of 0 or 1 and change from baseline ≤2 ▪ Observed score of 0	▪ Missing if baseline or observed value is missing. ▪ Single item, missing if missing.

Eczema Area and Severity Index (EASI)	The EASI assesses objective physician estimates of 2 dimensions of atopic dermatitis – disease extent and clinical signs (Hanifin et al 2001) – by scoring the extent of disease (percentage of skin affected: 0 = 0%; 1 = 1-9%; 2 = 10-29%; 3 = 30-49%; 4 = 50-69%; 5 = 70-89%; 6 = 90-100%) and the severity of 4 clinical signs (erythema, edema/papulation, excoriation, and lichenification) each on a scale of 0 to 3 (0 = none, absent; 1 = mild; 2 = moderate; 3 = severe) at 4 body sites (head and neck, trunk, upper limbs, and lower limbs). Half scores are allowed between severities 1, 2 and 3. Each body site will have a score that ranges from 0 to 72, and the final EASI score will be obtained by weight-averaging these 4 scores. Hence, the final EASI score will range from 0 to 72 for each time point.	EASI score	Derive EASI region score for each of head and neck, trunk, upper limbs, and lower limbs as follows: $EASI_{region}$ = (Erythema + edema/papulation + Excoriation + Lichenification) *(value from percentage involvement), where erythema, edema/papulation, excoriation, and lichenification are evaluated on a scale of 0 to 3 and value from percentage involvement is on a scale of 0 to 6. Then total EASI score is as follows: $EASI = 0.1*EASI_{head\ and\ neck} + 0.3*EASI_{trunk} + 0.2*EASI_{upper\ limbs} + 0.4*EASI_{lower\ limbs}$	N/A – partial assessments cannot be saved.
		▪ Change from baseline in EASI score ▪ Percent change from baseline EASI score	Change from baseline: observed EASI score – baseline EASI score % change from baseline: $100 \times \frac{Observed\ score - Baseline}{Baseline}$	Missing if baseline or observed value is missing.
		EASI50	% Improvement in EASI score from baseline ≥ 50%: % change from baseline ≤ -50	Missing if baseline or observed value is missing.
		EASI75	% Improvement in EASI score from baseline ≥75%: % change from baseline ≤-75	Missing if baseline or observed value is missing.
		EASI90	% Improvement in EASI score from baseline ≥90%: % change from baseline ≤-90	Missing if baseline or observed value is missing.
Body Surface Area (BSA) Affected by AD	Body surface area affected by AD will be assessed for 4 separate body regions and is collected as part of the EASI assessment: head and neck, trunk (including genital region), upper extremities, and lower extremities	BSA score	Use the percentage of skin affected for each region (0 to 100%) in EASI as follows: $BSA\ Total = 0.1*BSA_{head\ and\ neck} + 0.3*BSA_{trunk} + 0.2*BSA_{upper\ limbs} +$	N/A – partial assessments cannot be saved.

	(including the buttocks). Each body region will be assessed for disease extent ranging from 0% to 100% involvement. The overall total percentage will be reported based off of all 4 body regions combined, after applying specific multipliers to the different body regions to account for the percent of the total BSA represented by each of the 4 regions.		$0.4*BSA_{lower\ limbs}$	
		Change from baseline in BSA score	Change from baseline: observed BSA score – baseline BSA score	Missing if baseline or observed value is missing.
SCORing Atopic Dermatitis (SCORAD)	The SCORing Atopic Dermatitis (SCORAD) index uses the rule of nines to assess disease extent (head and neck 9%; upper limbs 9% each; lower limbs 18% each; anterior trunk 18%; back 18%; and genitals 1%). It evaluates 6 clinical characteristics to determine disease severity: (1) erythema, (2) edema/papulation, (3) oozing/crusts, (4) excoriation, (5) lichenification, and (6) dryness on a scale of 0 to 3 (0=absence, 1=mild, 2=moderate, 3=severe). The SCORAD index also assesses subjective symptoms of pruritus and sleep loss in the last 72 hours on visual analogue scales (VAS) of 0 to 10 where 0 is no itch or sleep loss and 10 is worst imaginable itch or sleep loss. These 3 aspects: extent of disease, disease severity, and subjective symptoms combine to give a maximum possible score of 103 (Stalder et al. 1993; Kunz et al. 1997; Schram et al. 2012).	SCORAD score	SCORAD = A/5 + 7B/2 + C, where A is extent of disease, range 0-100 B is disease severity, range 0-18 C is subjective symptoms, range 0-20	Missing if components A and B are missing or if component C is missing. Partial assessments performed by physician cannot be saved and partial assessments performed by subject cannot be saved.
		▪ Change from baseline in SCORAD score ▪ Percent change from baseline in SCORAD score	Change from baseline: observed SCORAD score – baseline SCORAD score % change from baseline: $100 \times \frac{Observed\ score - Baseline}{Baseline}$	Missing if baseline or observed value is missing.
		SCORAD75	% Improvement in SCORAD from baseline ≥75%: % change from baseline ≤-75	Missing if baseline or observed value is missing.
		SCORAD90	% Improvement in SCORAD from baseline ≥90%: % change from baseline ≤-90	Missing if baseline or observed value is missing.

Table JAIY.6.5. Description of Primary, Secondary and Exploratory Efficacy Analyses

Measure	Variable	Analysis Method (Section 6.2.3)	Population (Section 6.2.1)	Comparison/Time Point	Analysis Type
Validated Investigator's Global Assessment for AD (IGA)	Proportion of patients achieving IGA [0,1] with a ≥2-point improvement	Logistic regression using NRI	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Primary analysis
			PPS	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Sensitivity analysis
			ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 4	Secondary analysis
		Logistic regression using pMI and Tipping Point	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Sensitivity analysis
	Proportion of patients achieving IGA [0]	Logistic regression using NRI	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Secondary analysis
Eczema Area and Severity Index (EASI)	• EASI score • Change from baseline in EASI score • Percent change from baseline in EASI score	MMRM	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Key secondary analysis
			PPS	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Sensitivity analysis
		ANCOVA using mLOCF	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Sensitivity analysis
		pMI and Tipping Point	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Sensitivity analysis
	• Proportion of patients achieving EASI50	Logistic regression using NRI	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Secondary analysis
	• Proportion of patients achieving EASI75 • Proportion of patients achieving EASI90	Logistic regression using NRI	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Key secondary analysis
			PPS	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Sensitivity analysis
		pMI and Tipping Point	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Sensitivity analysis
Body Surface Area (BSA) Affected by AD	• BSA score • Change from baseline in BSA score	MMRM	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Secondary analysis
		ANCOVA using mLOCF	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Sensitivity analysis

Measure	Variable	Analysis Method (Section 6.2.3)	Population (Section 6.2.1)	Comparison/Time Point	Analysis Type
SCORing Atopic Dermatitis (SCORAD)	• SCORAD score • Change from baseline in SCORAD score • Percent change from baseline in SCORAD score	MMRM	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Secondary analysis
		ANCOVA using mLOCF	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Sensitivity analysis
	Proportion of patients achieving SCORAD75	Logistic regression using NRI	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Key secondary analysis
			PPS	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Sensitivity analysis
		Logistic regression using pMI	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Sensitivity analysis
	Proportion of patients achieving SCORAD90	Logistic regression using NRI	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Secondary analysis
		pMI and Tipping Point	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Sensitivity analysis
Skin Infections	Proportion of patients developing skin infections requiring antibiotic treatment	Fisher's exact	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Secondary analysis

Abbreviations: ANCOVA = analysis of covariance; Bari = baricitinib; ITT = intent-to-treat; mLOCF = modified last observation carried forward; MMRM = mixed model repeated measures; NRI = nonresponder imputation; PBO = placebo; pMI=placebo multiple imputation; PPS = per protocol set.

Notes: (1) for all other post-baseline visits not mentioned in the table, but collected for the measures as specified in the protocol, the analyses will be made as exploratory analyses.

(2) All primary and key secondary analyses will be performed for Japan population. Other key secondary and exploratory analysis may be performed for Japan population.

6.12.1. Primary Outcome and Methodology

The validated Investigator's Global Assessment for AD (IGA) uses the clinical characteristics of erythema, papulation/induration, oozing/crusting and lichenification to produce a single-item score ranging from 0 to 4. The primary analysis of the study is to test the null hypotheses that neither baricitinib 4 mg nor baricitinib 2 mg is superior to placebo when evaluating the proportion of patients achieving IGA of 0 or 1 at Week 16 in the ITT population. The analysis assumes that treatment response disappears after patients are rescued or permanently discontinue from treatment. This will serve as the primary estimand. In this estimand, missing data due to the application of the primary censoring rule and the occurrence of other non-censor intercurrent events will be imputed using the NRI method described in Section 6.4.1.

A supplemental estimand is to test the null hypotheses that neither baricitinib 4 mg nor baricitinib 2 mg is superior to placebo when evaluating the proportion of patients achieving IGA of 0 or 1 at Week 16 in the ITT population. This analysis assumes the treatment response disappears after patients permanently discontinue from treatment. In this supplemental estimand, missing data due to the application of the secondary censoring rule and the occurrence of other non-censor intercurrent events will be imputed using the NRI method described in Section 6.4.1.

A logistic regression analysis as described in Section 6.2.3 will be used for the comparisons. The odds ratio, the corresponding 95% CIs and p-value, as well as the treatment differences and the corresponding 95% CIs, will be reported.

Multiplicity controlled analyses will be performed on the primary and key secondary (see Section 4.2.1) objectives to control the overall Type I error rate at a 2-sided alpha level of 0.05. A graphical approach will be used to perform the multiplicity controlled analyses as described in Section 6.6.

6.12.2. Secondary and Exploratory Efficacy Analyses

For secondary analysis, the null hypotheses is that neither baricitinib 4 mg nor baricitinib 2 mg is superior to placebo in the ITT population. These analyses assume treatment response disappears after patients are rescued or permanently discontinued from treatment and will serve as the primary estimand. In this estimand, missing data due to the application of the primary censoring rule and the occurrence of other non-censor intercurrent events will be imputed using the method described in Table JAIY.6.1.

A supplemental estimand for secondary endpoints is to test the null hypotheses that neither baricitinib 4 mg nor baricitinib 2 mg is superior to placebo in the ITT population. These analyses assume the treatment response disappears after patients permanently discontinue from treatment. In this supplemental estimand, missing data due to the application of the secondary censoring rule and the occurrence of other non-censor intercurrent events will be imputed using the method described in Table JAIY.6.1.

A list of exploratory endpoints are provided in Section 4.2.2. There will be no adjustment for multiple comparisons for exploratory endpoints. The secondary and exploratory efficacy

endpoints are detailed in Table JAIY.6.4 and analyses are provided in Table JAIY.6.5. Health outcomes analyses are described in Section 6.13.

6.12.3. Sensitivity Analyses

Sensitivity analyses are included to demonstrate robustness of analyses methods using different missing data imputations, censoring rules, populations and analyses assumptions. Sensitivity analyses for select outcomes have been previously described and include the following:

- Analyses of key endpoints using the per-protocol analysis set (Section 6.2.1)
- Analyses of key endpoints using the secondary censoring rule (Section 6.2)
- Placebo multiple imputation (Section 6.4.4)
- Tipping point analysis (Section 6.4.5)
- The addition of a treatment-by-region interaction to the logistic regression model for the primary outcome (Section 6.5)
- Analysis of continuous outcomes with ANCOVA (Section 6.2.3), with missing data imputed using mLOCF (Section 6.4.3).

6.13. Health Outcomes/Quality-of-Life Analyses

The general methods used to summarize health outcomes and quality-of-life measures, including the definition of baseline value for assessments are described in Section 6.1.

Health outcomes and quality-of-life measures will generally be analyzed according to the formats discussed in Section 6.12.

Table JAIY.6.6 includes the descriptions and derivations of the health outcomes and quality-of-life measures.

Table JAIY.6.7 provides the detailed analyses including analysis type, method and imputation, population, time point, and comparisons for health outcomes and quality-of-life measures.

Table JAIY.6.6. Description and Derivation of Health Outcomes and Quality-of-Life Measures

Measure	Description	Variable	Derivation / Comment	Imputation Approach if Missing Components
Itch Numeric Rating Scale (NRS)	The Itch Numeric Rating Scale (NRS) is a patient-administered, 11-point horizontal scale anchored at 0 and 10, with 0 representing "no itch" and 10 representing "worst itch imaginable." Overall severity of a patient's itching is indicated by selecting the number that best describes the worst level of itching in the past 24 hours (Naegeli et al. 2015; Kimball et al. 2016). Refer to Section 6.2.2 for details on how to calculate the weekly score which will be used in the continuous analysis.	Itch NRS score	Single item; range 0-10. Refer to Section 6.2.2 on how to derive the visit score.	Refer to Section 6.2.2 on how to derive the weekly visit score.
		▪ Change from baseline in Itch NRS ▪ Percent change from baseline in Itch NRS	Change from baseline: observed Itch score – baseline Itch score % change from baseline: $100 \times \frac{Observed\ score - Baseline}{Baseline}$	Missing if baseline or observed value is missing.
		4-point Itch improvement in subgroup of patients with baseline Itch NRS ≥4	Change from baseline ≤-4 and baseline ≥4	Missing if baseline is missing or <4 or observed value is missing.
		Itch-free days (Itch NRS = 0)	The number of itch-free days during intervals starting on the day of the first study drug administration. This will be calculated for the following intervals: baseline to Week 4, Week 4 to Week 8, Week 8 to Week 12 and Week 12 to Week 16. Day 1 is defined as the day of first study drug administration therefore the baseline to Week 4 assessment is based on Day 1 to Day 28, Week 4 to Week 8 is based on Day 29 to Day 56, etc.	Missing if observed value is missing.

Measure	Description	Variable	Derivation / Comment	Imputation Approach if Missing Components
Skin Pain Numeric Rating Scale (NRS)	Skin Pain NRS is a patient-administered, 11-point horizontal scale anchored at 0 and 10, with 0 representing "no pain" and 10 representing "worst pain imaginable." Overall severity of a patient's skin pain is indicated by selecting the number that best describes the worst level of skin pain in the past 24 hours Refer to Section 6.2.2 for details on how to calculate the weekly score which will be used in the continuous analysis.	Skin Pain NRS score	Single item; range 0 to 10. Refer to Section 6.2.2 on how to derive the visit score.	Refer to Section 6.2.2 on how to derive the visit score.
		Change from baseline in Skin Pain NRS	Change from baseline: observed skin pain score – baseline skin pain score	Missing if baseline or observed value is missing.
		Skin Pain-free days (Skin Pain NRS = 0)	The number of skin pain-free days during intervals starting on the day of the first study drug administration. This will be calculated for the following intervals: baseline to Week 4, Week 4 to Week 8, Week 8 to Week 12 and Week 12 to Week 16. Thus, if Day 1 is defined as the day of first study drug administration, the baseline to Week 4 assessment is based on Day 1 to Day 28, Week 4 to Week 8 is based on Day 29 to Day 56, etc.	Missing if observed value is missing.

Measure	Description	Variable	Derivation / Comment	Imputation Approach if Missing Components
Atopic Dermatitis Sleep Scale (ADSS)	The Atopic Dermatitis Sleep Scale (ADSS) is a 3-item, patient-administered questionnaire developed to assess the impact of itch on sleep including difficulty falling asleep, frequency of waking, and difficulty getting back to sleep last night. Patient's rate their difficulty falling asleep and difficulty getting back to sleep, items 1 and 3, respectively, using a 5-point Likert-type scale with response options ranging from 0 "not at all" to 4 "very difficult." Patients report their frequency of waking last night, item 2, by selecting the number of times they woke up each night, ranging from 0 to 29 times. The ADSS is designed to be completed each day with respondents thinking about sleep "last night." Each item is scored individually.	▪ Item 1 score of ADSS ▪ Item 2 score of ADSS ▪ Item 3 score of ADSS	Single items: Item 1, range 0 to 4; Item 2, range 0 to 29; Item 3, range 0 to 4. Refer to Section 6.2.2 on how to derive the visit score.	Refer to Section 6.2.2 on how to derive the weekly visit score.
		▪ Change from baseline in score of Item 1 of ADSS ▪ Change from baseline in score of Item 2 of ADSS ▪ Change from baseline in score of Item 3 of ADSS	Change from baseline: observed ADSS item score – baseline ADSS item score	Missing if baseline or observed value is missing.

Measure	Description	Variable	Derivation / Comment	Imputation Approach if Missing Components
Patient- Oriented Eczema Measure (POEM)	The POEM is a simple, 7-item, patient-administered scale that assesses disease severity in children and adults. Patients respond to questions about the frequency of 7 symptoms (itching, sleep disturbance, bleeding, weeping/oozing, cracking, flaking, and dryness/roughness) over the last week. Response categories include "No days," "1-2 days," "3-4 days," "5-6 days," and "Every day" with corresponding scores of 0, 1, 2, 3, and 4, respectively. Scores range from 0-28 with higher total scores indicating greater disease severity (Charman et al. 2004).	POEM score	POEM total score: sum of questions 1 to 7, Range 0 to 28.	If a single question is left unanswered, then that question is scored as 0. If more than one question is unanswered, then the tool is not scored. If more than one response is selected, then the response with the highest score is used.
		Change from baseline in POEM score	Change from baseline: observed POEM score – baseline POEM score	Missing if baseline or observed value is missing.
		4-point improvement in POEM score in subgroup of patients with baseline ≥4	Change from baseline ≤-4 and baseline ≥4	Missing if baseline is missing or <4 or observed value is missing.

Measure	Description	Variable	Derivation / Comment	Imputation Approach if Missing Components
Patient Global Impression of Severity–Atopic Dermatitis (PGI-S-AD)	The Patient Global Impression of Severity–Atopic Dermatitis (PGI-S-AD) is a single-item question asking the patient how they would rate their overall AD symptoms over the past 24 hours. The 5 categories of responses range from "no symptoms" to "severe."	PGI-S-AD score	Single item. Range 1 to 5. Refer to Section 6.2.2 on how to derive the visit score.	Refer to Section 6.2.2 on how to derive the visit score.
		Change from baseline in PGI-S-AD	Change from baseline: observed PGI-S-AD score – baseline PGI-S-AD score	Missing if baseline or observed value is missing.
Hospital Anxiety Depression Scale (HADS)	The Hospital Anxiety Depression Scale (HADS) is a 14-item self-assessment scale that determines the levels of anxiety and depression that a patient is experiencing over the past week. The HADS utilizes a 4-point Likert scale (e.g., 0 to 3) for each question and is intended for ages 12 to 65 years (Zigmond and Snaith 1983; White et al. 1999). Scores for each domain (anxiety and depression) can range from 0 to 21, with higher scores indicating greater anxiety or depression (Zigmond and Snaith 1983; Snaith 2003).	HADS score for anxiety and depression domains	Anxiety domain score is sum of the seven anxiety questions, range 0 to 21; Depression domain score is sum of the seven depression questions, range 0 to 21.	N/A – partial assessments cannot be saved.
		Change from baseline in HADS total score, anxiety and depression domain	Change from baseline: observed HADS domain score – baseline HADS domain score	Missing if baseline or observed value is missing.
Dermatology Life Quality Index (DLQI)	The Dermatology Life Quality Index (DLQI) is a simple, patient-administered, 10-item, validated, quality-of-life questionnaire that covers 6 domains including symptoms and feelings, daily activities, leisure, work and school, personal relationships, and treatment. The recall period of this scale is over the "last week." Response categories include "a little," "a lot," and "very much," with corresponding	Symptoms and feelings domain	Sum of questions 1 and 2, range 0 to 6.	N/A – partial assessments cannot be saved.
		Daily activities domain	Sum of questions 3 and 4, range 0 to 6.	N/A – partial assessments cannot be saved.
		Leisure domain	Sum of questions 5 and 6, range 0 to 6.	N/A – partial assessments cannot be saved.

Measure	Description	Variable	Derivation / Comment	Imputation Approach if Missing Components
	scores of 1, 2, and 3, respectively, and "not at all," or unanswered ("not relevant") responses scored as 0. Scores range from 0-30 with higher scores indicating greater impairment of quality of life. A DLQI total score of 0 to 1 is considered as having no effect on a patient's health-related QoL (Hongbo et al. 2005), and a 4-point change from baseline is considered as the minimal clinically important difference threshold (Khilji et al. 2002; Basra et al. 2015).	Work and school domain	Sum of questions 7 and 7B (if it is answered), range 0 to 3. Responses of "yes" and "no" on Question 7 are given scores of 3 and 0 respectively. If Question 7 is answered "no" then Question 7b is answered with "a lot", "a little", "not at all" getting scores of 2, 1, 0 respectively.	N/A – partial assessments cannot be saved.
		Personal relationships domain	Sum of questions 8 and 9, range 0 to 6.	N/A – partial assessments cannot be saved.
		Treatment domain	Question 10, range 0 to 3.	N/A – partial assessments cannot be saved.
		DLQI total score	DLQI total score: sum of all six DLQI domain scores, range 0 to 30.	N/A – partial assessments cannot be saved.
		Change from baseline in DLQI	Change from baseline: observed DLQI score – baseline DLQI score	Missing if baseline or observed value is missing.
		4-point improvement in DLQI total score in subgroup of patients with baseline ≥4	Change from baseline ≤4 and baseline ≥4	Missing if baseline is missing or <4 or observed value is missing.
		DLQI (0,1)	A DLQI (0,1) response is defined as a post-baseline DLQI total score of 0 or 1	Missing if the DLQI total score is missing

Measure	Description	Variable	Derivation / Comment	Imputation Approach if Missing Components
Work Productivity and Activity Impairment: Atopic Dermatitis (WPAI-AD)	The Work Productivity and Activity Impairment Questionnaire–Atopic Dermatitis (WPAI-AD) records impairment due to AD during the past 7 days. The WPAI-AD consists of 6 items grouped into 4 domains: absenteeism (work time missed), presenteeism (impairment at work/reduced on-the-job effectiveness), work productivity loss (overall work impairment/absenteeism plus presenteeism), and activity impairment. Scores are calculated as impairment percentages (Reilly et al. 1993), with higher scores indicating greater impairment and less productivity.	Employment status	Question (Q)1	Single item, missing if missing.
		Change in employment status	Employed at baseline and remained employed: Q1 = 1 at post-baseline visit and at baseline visit. Not employed at baseline and remain unemployed: Q1 = 0 at post-baseline visit and at baseline visit.	Missing if baseline or observed value is missing.
		Percentage of absenteeism	Percent work time missed due to problem: (Q2/(Q2 + Q4))*100	If Q2 or Q4 is missing, then missing.
		Change from baseline in absenteeism	Change from baseline: observed absenteeism – baseline absenteeism	Missing if baseline or observed value is missing.
		Percentage of presenteeism	Percent impairment (reduced productivity while at work) while working due to problem: (Q5/10)*100	If Q5 is missing, then missing.
		Change from baseline in presenteeism	Change from baseline: observed presenteeism – baseline absenteeism	Missing if baseline or observed value is missing.
		Overall work impairment	Percent overall work impairment (combines absenteeism and presenteeism) due to problem: (Q2/(Q2+Q4) + [(1-Q2/(Q2+Q4))*(Q5/10)])*100	If Q2, Q4, or Q5 is missing, then missing.

Measure	Description	Variable	Derivation / Comment	Imputation Approach if Missing Components
		Change from baseline in work impairment	Change from baseline: observed work impairment – baseline work impairment	Missing if baseline or observed value is missing.
		Percentage of impairment in activities	Percent activity impairment (performed outside of work) due to problem: (Q6/10)*100	If Q6 is missing, then missing.
		Change from baseline in impairment in activities	Change from baseline: observed impairment in activities – baseline impairment in activities	Missing if baseline or observed value is missing.
European Quality of Life–5 Dimensions–5 Levels (EQ-5D-5L)	The European Quality of Life–5 Dimensions–5 Levels (EQ-5D-5L) is a standardized measure of health status that provides a simple, generic measure of health for clinical and economic appraisal. The EQ-5D-5L consists of 2 components: a descriptive system of the respondent's health and a rating of his or her current health state using a 0 to 100 mm VAS. The descriptive system comprises the following 5 dimensions: mobility, self-care, usual activities, pain/discomfort, and anxiety/depression. Each dimension has 5 levels: no problems, slight problems, moderate problems, severe problems, and extreme problems. The respondent is asked to indicate his or her health state by ticking (or placing a cross) in the box associated with the most appropriate statement in each of the 5 dimensions. It should be noted that the numerals 1 to 5 have no arithmetic properties and should not be used as an ordinal score. The VAS records the respondent's self-rated health on a vertical VAS where the endpoints are	▪ EQ-5D mobility ▪ EQ-5D self-care ▪ EQ-5D usual activities ▪ EQ-5D pain/ discomfort ▪ EQ-5D anxiety/ depression	Five health profile dimensions, each dimension has 5 levels: 1 = no problems 2 = slight problems 3 = moderate problems 4 = severe problems 5 = extreme problems It should be noted that the numerals 1 to 5 have no arithmetic properties and should not be used as a primary score.	Each dimension is a single item, missing if missing.
		EQ-5D VAS	Single item. Range 0 to 100. 0 represents "worst health you can imagine" 100 represents "best health you can imagine"	Single item, missing if missing.
		Change from baseline in EQ-5D VAS	Change from baseline: observed EQ-5D VAS score – baseline EQ-5D VAS score	Missing if baseline or observed value is missing.
		EQ-5D-5L UK Population-based index score (health	Derive EQ-5D-5L UK Population-based index score according to the link by using the	N/A – partial assessments cannot be saved

Measure	Description	Variable	Derivation / Comment	Imputation Approach if Missing Components
	labeled "best imaginable health state" and "worst imaginable health state." This information can be used as a quantitative measure of health outcome. The EQ-5D-5L health states, defined by the EQ-5D-5L descriptive system, may be converted into a single summary index by applying a formula that essentially attaches values (also called weights) to each of the levels in each dimension (Herdman et al. 2011; EuroQol Group 2015 [WWW]).	state index)	UK algorithm to produce a patient-level index score between -0.59 and 1.0 (continuous variable).	on the eCOA tablet.
		Change from baseline in EQ-5D-5L UK Population-based index score	Change from baseline: observed EQ-5D-5L UK score – baseline EQ-5D-5L UK score	Missing if baseline or observed value is missing.
		EQ-5D-5L US Population-based index score (health state index)	Derive EQ-5D-5L US Population-based index score according to the link by using the US algorithm to produce a patient-level index score between -0.11 and 1.0 (continuous variable).	N/A – partial assessments cannot be saved on the eCOA tablet.
		Change from baseline in EQ-5D-5L US Population-based index score	Change from baseline: observed EQ-5D-5L US score – baseline EQ-5D-5L US score	Missing if baseline or observed value is missing.
PROMIS Itch Questionnaire (PIQ)	PIQ – Itch Interference: consists of 8 items assessing the impact of itch on various aspects of life. PIQ – Activity and Clothing: consists of 8 items assessing activity and clothing related quality of life impairment from itch in adults "in the past 7 days". PIQ – Mood and Sleep: consists of 8 items assessing mood and sleep related quality of life impairment from itch and impact of itch "in the past 7 days". PIQ – Scratching Behavior: consists of 5 items assessing quality of life impairment from scratching behavior and the physical	PIQ – Itch Interference[a]	8 items. Each range 1 to 5. 1=Never, 2=Rarely, 3=Sometimes, 4=Often, 5=Almost always (continuous variable). Total raw scores are converted to T-Scores with higher scores representing greater impact because of itch.	Calculation is made by HealthMeasures Scoring Service, powered by Assessment Center[SM]
		Change from baseline in PIQ – Itch Interference	Change from baseline: observed PIQ Itch Interference score – baseline PIQ Itch Interference score	Missing if Baseline or observed value is missing.
		PIQ – Activity and Clothing	8 items. Each range 1 to 5. 1=Never, 2=Rarely,	Score is calculated by

Measure	Description	Variable	Derivation / Comment	Imputation Approach if Missing Components
	manifestations of itch in adults "in the past 7 days".		3=Sometimes, 4=Often, 5=Almost always (continuous variable). Total raw scores are converted to T-Scores with higher scores representing greater impact because of itch.	HealthMeasures Scoring Service, powered by Assessment Center[SM]
		Change from baseline in PIQ – Activity and Clothing	Change from baseline: observed PIQ Activity and Clothing score – baseline PIQ Activity and Clothing score	Missing if Baseline or observed value is missing.
		PIQ – Mood and Sleep	8 items. Each range 1 to 5. 1=Never, 2=Rarely, 3=Sometimes, 4=Often, 5=Almost always (continuous variable). Total raw scores are converted to T-Scores with higher scores representing greater impact because of itch.	Score is calculated by HealthMeasures Scoring Service, powered by Assessment Center[SM]
		Change from baseline in PIQ – Mood and Sleep	Change from baseline: observed PIQ Mood and Sleep score – baseline PIQ Mood and Sleep score	Missing if Baseline or observed value is missing.
		PIQ – Scratching Behavior	5 items. Each range 1 to 5. Response options for the frequency of scratching behaviors: 1=Never, 2=Rarely, 3=Sometimes, 4=Often, 5=Almost always The response options for the worry related to scratching items: 1=Not at all, 2=A little bit, 3=Somewhat, 4=Quite a bit, 5=Very much (continuous variable). Total raw scores are converted to T-Scores with higher	Score is calculated by HealthMeasures Scoring Service, powered by Assessment Center[SM]

Measure	Description	Variable	Derivation / Comment	Imputation Approach if Missing Components
			scores representing more scratching behavior.	
		Change from baseline in PIQ – Scratching Behavior	Change from baseline: observed PIQ Scratching Behavior score – baseline PIQ Scratching Behavior score	Missing if Baseline or observed value is missing.
PROMIS – Sleep Related Impairment	The Sleep Related Impairment Short Form within the PROMIS bank consists of 8 items measuring self-reported perceptions of alertness, sleepiness, and tiredness during usual waking hours, and the perceived functional impairments during wakefulness associated with sleep problems or impaired alertness "in the past 7 days". Response options range from 1=Not at all; 2=A little bit; 3=Somewhat; 4=Quite a bit; to 5=Very much.	PROMIS sleep related impairment	8 items. Each range 1 to 5. 1=Not at all; 2=A little bit; 3=Somewhat; 4=Quite a bit; to 5=Very much (continuous variable). Total raw scores are converted to T-Scores with higher scores representing greater sleep impairment.	Score is calculated by HealthMeasures Scoring Service, powered by Assessment Center[SM]
		Change from baseline in PROMIS sleep related impairment	Change from baseline: observed PROMIS Sleep Related Impairment score – baseline PROMIS Sleep Related Impairment score	Missing if Baseline or observed value is missing.
Neuro-QoL – Cognitive Function	The Cognitive Function Short Form domain within Neuro-QoL bank consists of 8-items measuring Executive Function (perceived difficulties in applications of mental health function related to planning, organizing, calculating, remembering and learning) "in the past 7 days" and General Concerns (perceived difficulties in everyday cognitive abilities such as memory, attention, and decision making) using the lead-in phrase "how much difficulty do you currently have…". The response options for the Executive Function items range from 1=Very Often (several times a day); 2=Often (once a day); 3=Sometimes (2-3 times); 4=Rarely (once); to	Neuro-QoL – Cognitive Function	The total raw scores are converted to T-Scores with higher scores indicating better (desirable) self-reported health.	Score is calculated by HealthMeasures Scoring Service, powered by Assessment Center[SM]
		Change from baseline in Neuro-QoL – Cognitive Function	Change from baseline: observed Neuro-QoL – Cognitive Function score – baseline Neuro-QoL – Cognitive Function score	Missing if Baseline or observed value is missing.

Measure	Description	Variable	Derivation / Comment	Imputation Approach if Missing Components
	5=Never). The response options for the General Concerns items range from 1=Cannot do; 2=A lot; 3=Somewhat; 4=A little, to 5=None.			
Patient Benefit Index (PBI)	The Patient Benefit Index (PBI) measures patient-defined treatment objectives and benefits. It consists of 2 questionnaires. Before therapy, patients complete the standardized "Patient Needs Questionnaire" (PNQ) indicating individual importance of treatment objectives. This reflects their personal preferences with respect to therapeutic benefit. During the study patients rate the extent to which the treatment objectives have been achieved in the "Patient Benefit Questionnaire" (PBQ). Subscales of the Patient Benefit Index are: • Reducing social impairments subscale score: item 11, 13, 14, 15, 16, 17 • Reducing psychological impairments subscale score: item 6, 7, 9, 10, 12 • Reducing impairments due to therapy subscale score: item 18, 19, 20, 21 • Reducing physical impairments subscale score: item 1, 2, 3, 4, 5 • Having confidence in healing subscale score: item 8, 22, 23	• Patient Benefit Index (PBI) global score • Subscale scores	25 items. 0=not at all, 1=somewhat; 2=moderately; 3=quite; 4=very; PBI global score is calculated for each patient by weighing the achievement values of the treatment objectives by their importance to the individual patient. $$PBI = \sum_{i=1}^{k} \frac{PNQ_i}{\sum_{i=1}^{k} PNQ_i} PBQ_i$$ For score calculation, both "does/did not apply" and question unanswered will be treated as missing values. The global score is calculated using only these item pairs, for which the patient has given a response other than "does/did not apply to me" in both PNQ and PBQ. Subscale scores are calculated in the same manner as the global score.	PBI global score and subscales may only be computed if the patient has provided at least 75% valid data in each of the PNQ and PBQ respectively. In this context, the responses "not at all" and "does/did not apply" count as valid data. Thus a treatment goal is regarded missing if the patient has not responded to the item in the PNQ and/or in the PBI.

Measure	Description	Variable	Derivation / Comment	Imputation Approach if Missing Components
		• PBI global score ≥1	PBI global score at least 1	Missing if observed value is missing

[a] PIQ – Itch Interference here is the same as PIQ – General in the protocol.

Table JAIY.6.7. Description of Health Outcomes and Quality-of-Life Measures Analyses

Measure	Variable	Analysis Method (Section 6.2.3)	Population (Section 6.2.1)	Comparison/Time Point	Analysis Type
Itch Numeric Rating Scale (NRS)	• Itch NRS score • Change from baseline in Itch NRS score	MMRM	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 1, …, 16	Secondary Analysis
		ANCOVA using mLOCF	ITT	Bari 4 mg or Bari 2 mg vs PBO; Day 2, Week 1,…, 16	Secondary Analysis at Day 2; Sensitivity Analysis for others
		pMI	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 1, …, 16	Sensitivity Analysis
	Percent change from baseline Itch score	MMRM	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 1, …,16	Secondary Analysis
		ANCOVA using mLOCF	ITT	Bari 4 mg or Bari 2 mg vs PBO; Day 2, Week 1, …, 16	Secondary Analysis at Day 2; Sensitivity Analysis for others
	Proportion of patients achieving a 4-point improvement in Itch NRS in subgroup of patients who had baseline Itch NRS ≥4	Logistic regression using NRI	ITT	Bari 4 mg or Bari 2 mg vs PBO; Day 2, Week 1,…, 16	Key Secondary Analysis at Day 2, Week 1, 2 and 16
			PPS	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Sensitivity analysis
		Logistic regression using pMI and Tipping Point	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Sensitivity analysis
	Number of Itch-free (Itch NRS = 0) Days	Two-sample t-test	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 0 to Week 4, …, Week 12 to 16	Exploratory Analysis
Skin Pain Numeric Rating Scale (NRS)	• Skin Pain NRS score • Change from baseline in Skin Pain NRS score	MMRM	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 1,…, 16	Key Secondary Analysis at Week 16
			PPS	Bari 4 mg or Bari 2 mg vs PBO; Week 1,…, 16	Sensitivity analysis

Measure	Variable	Analysis Method (Section 6.2.3)	Population (Section 6.2.1)	Comparison/Time Point	Analysis Type
		ANCOVA using mLOCF	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 1,…, 16	Sensitivity Analysis
		pMI	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 1,…, 16	Sensitivity analysis
	Number of Skin Pain-free (Skin pain NRS = 0) Days	Two-sample t-test	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 0 to Week 4,…, Week 12 to 16	Exploratory Analysis
	Proportion of patients achieving a 4-point improvement in Skin Pain NRS in subgroup of patients with baseline Skin Pain NRS ≥4	Logistic regression using NRI	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 1,…, 16	Other Secondary Analysis
Atopic Dermatitis Sleep Scale (ADSS)	• ADSS item scores • Change from baseline in ADSS item scores	MMRM	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 1,…, 16	Key Secondary Analysis at Week 1 and 16
			PPS	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Sensitivity analysis
		ANCOVA using mLOCF	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 1,…, 16	Sensitivity Analysis
		pMI	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Sensitivity analysis
Patient-Oriented Eczema Measure (POEM)	• POEM score • Change from baseline in POEM score	MMRM	ITT	Bari 4 mg or Bari 2 mg vs PBO; at each post-baseline visit	Secondary Analysis
		ANCOVA using mLOCF	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Sensitivity Analysis
	Proportion of patients achieving a 4-point improvement in subgroup of patients with baseline POEM ≥ 4	Logistic regression using NRI	ITT	Bari 4 mg or Bari 2 mg vs PBO; at each post-baseline visit	Secondary Analysis
Patient Global Impression of Severity–Atopic Dermatitis (PGI-S-AD)	• PGI-S-AD score • Change from baseline in PGI-S-AD score	MMRM	ITT	Bari 4 mg or Bari 2 mg vs PBO; at each post-baseline visit	Secondary Analysis
		ANCOVA using mLOCF	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Sensitivity Analysis

Measure	Variable	Analysis Method (Section 6.2.3)	Population (Section 6.2.1)	Comparison/Time Point	Analysis Type
Hospital Anxiety Depression Scale (HADS)	• HADS domain scores • Change from baseline in HADS domain	MMRM	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 2, 4, 8,16	Secondary Analysis
		ANCOVA using mLOCF	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Sensitivity Analysis
Dermatology Life Quality Index (DLQI)	• DLQI total score • Change from baseline in DLQI • Observed and change from baseline in domain scores -Symptoms and feelings -Daily activities -Leisure -Work and school -Personal relationships -Treatment	MMRM	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Secondary Analysis
		ANCOVA using mLOCF	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Sensitivity Analysis
	Proportion of patients achieving a 4-point improvement in DLQI total score in subgroup of patients with baseline DLQI ≥4	Logistic regression using NRI	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 1, 2, 4, 8, 16	Other Secondary Analysis
	Proportion of patients achieving DLQI (0,1)	Logistic regression using NRI	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 1, 2, 4, 8, 16	Other Secondary Analysis
Work Productivity and Activity Impairment: Atopic Dermatitis (WPAI-AD)	Observed and Change from baseline in employment status	Descriptive statistics (observed)	ITT	No comparisons; Week 16	Secondary Analysis
	Observed and Change from baseline in: • absenteeism • presenteeism • overall work impairment • impairment in activities	MMRM	ITT	Bari 4 mg or Bari 2 mg vs PBO; at each post-baseline visit	Secondary Analysis
		ANCOVA using mLOCF	ITT	Bari 4 mg or Bari 2 mg vs PBO; at each post-baseline visit	Sensitivity Analysis

Measure	Variable	Analysis Method (Section 6.2.3)	Population (Section 6.2.1)	Comparison/Time Point	Analysis Type
European Quality of Life–5 Dimensions–5 Levels (EQ-5D-5L)	Observed values in • EQ-5D mobility • EQ-5D self-care • EQ-5D usual activities • EQ-5D pain/ discomfort • EQ-5D anxiety/ depression	Logistic Regression using NRI	ITT	Bari 4 mg or Bari 2 mg vs PBO: at each post-baseline visit	Exploratory Analysis
	Observed and Change from baseline in • EQ-5D VAS • EQ-5D-5L UK Population-based index score • EQ-5D-5L US Population-based index score	MMRM	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Secondary Analysis
		ANCOVA using mLOCF	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Sensitivity Analysis
PROMIS	Observed and Change from baseline in: • PIQ – Itch Interference score • PIQ – Activity and Clothing score • PIQ – Mood and Sleep score • PIQ – Scratching Behavior score • PROMIS – Sleep-Related Impairment score	MMRM	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 1, 2, 4, 8, 12, 16	Exploratory Analysis
		ANCOVA using mLOCF	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 1, 2, 4, 8, 12, 16	Sensitivity Analysis
Neuro-QoL – Cognitive Function score	Observed and Change from baseline in: Neuro-QoL – Cognitive Function score	ANCOVA using mLOCF	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Exploratory Analysis

Measure	Variable	Analysis Method (Section 6.2.3)	Population (Section 6.2.1)	Comparison/Time Point	Analysis Type
PBI	• PBI global score • Reducing social impairments subscale score • Reducing psychological impairments subscale score • Reducing impairments due to therapy subscale score • Reducing physical impairments subscale score • Having confidence in healing subscale score	ANOVA[a]	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Exploratory Analysis
PBI	Proportion of patients achieving PBI global score ≥1 at Week 16	Logistic regression using NRI	ITT	Bari 4 mg or Bari 2 mg vs PBO; Week 16	Exploratory Analysis

Abbreviations: ANCOVA = analysis of covariance; ANOVA= analysis of variance; Bari = baricitinib; ITT = intent-to-treat; mLOCF = modified last observation carried forward; MMRM = mixed model repeated measures; NRI = nonresponder imputation; PBO = placebo; pMI=placebo multiple imputation; PPS = per protocol set.

Notes: for all other post-baseline visits not mentioned in the table, but collected for the measures as specified in the protocol, the analyses are made as exploratory analyses.

Notes: all the key secondary are performed for Japan population. Other secondary and exploratory analyses may be performed for Japan population.

[a] ANOVA model includes region, baseline disease severity (IGA) and treatment as factors in the model.

6.14. Bioanalytical and Pharmacokinetic/Pharmacodynamic Methods

Pharmacokinetic (PK), Pharmacodynamic (PD) and Biomarker analyses to address secondary and exploratory objectives of this study will be described by Lilly in separate PK/PD and Biomarker analysis plans.

6.15. Safety Analyses

The general methods used to summarize safety data, including the definition of baseline and postbaseline are described in Section 6.2.

Safety analyses will include data from first dose of the study treatment to after rescue, unless otherwise stated, and patients will be analyzed according to the investigational product to which they were randomized at Visit 2. A sensitivity approach to the safety analyses will use data censored at last dose of the study drug for patients rescued to systemic therapy. These analyses will be conducted for treatment-emergent adverse events (TEAEs), serious adverse events (SAEs), AEs leading to permanent study drug discontinuation and special topics excluding deaths and malignancies. Additional analyses may be conducted using data after rescue to systemic therapy for some safety topics such as systemic TEAEs, and SAEs. Safety analyses will use the safety population defined in Section 6.2.1.

Safety topics that will be addressed include the following: AEs including TEAEs and SAEs, clinical laboratory evaluations, vital signs and physical characteristics, Columbia Suicide Severity Rating Scale (C-SSRS), the Self-Harm Supplement Form, safety in special groups and circumstances, including adverse events of special interest (AESI) (see Section 6.15.5), and investigational product interruptions.

Unless otherwise specified, by-visit summaries will include planned on-treatment visits. For tables that summarize events (such as AEs, categorical lab abnormalities, shift to maximum value), post-last dose follow-up data will be included. Follow-up data is defined as all data occurring up to 30 days (planned maximum follow-up time) after last dose of treatment including rescue, regardless of study period.

For selected safety assessments other than events, descriptive statistics may be presented for the last measure observed during post-treatment follow-up (up to 30 days after the last dose of treatment including rescue, regardless of study period).

6.15.1. Extent of Exposure

Duration of exposure (in days) will be calculated as follows:

- Duration of exposure to investigational product (including exposure after the initiation of rescue therapy): *date of last dose of study drug including rescue – date of first dose of study drug + 1*.

Last dose of study drug including rescue is calculated as last date on study drug. See the compound level safety standards for more details.

Total patient-years (PY) of exposure to study drug will be reported for each treatment group for overall duration of exposure. Descriptive statistics will be provided for patient-days of exposure and the frequency of patients falling into different exposure ranges in addition to cumulative exposures will be summarized.

Exposure ranges will be summarized as follows:

- ≥28 days, ≥56 days, ≥84 days, and ≥112 days
- >0 to <28 days, ≥28 days to <56 days, ≥56 days to <84 days, ≥84 days to <112 days, and ≥112 days

Overall exposure for a treatment group will be summarized in total PY which is calculated according to the following formula:

- *Exposure in PY (PYE) = sum of duration of exposure in days* (for all patients in treatment group) */ 365.25*

6.15.2. Adverse Events

Adverse events are recorded in the eCRFs. Each AE will be coded to system organ class (SOC) and preferred term (PT) using the Medical Dictionary for Regulatory Activities (MedDRA) version that is current at the time of database lock. Severity of AEs is recorded as mild, moderate, or severe.

A TEAE is defined as an event that either first occurred or worsened in severity after the first dose of study treatment and on or prior to the last visit date during the analysis period. The analysis period is defined as the treatment period plus up to 30 days off-drug follow-up time.

Adverse events are classified based upon the MedDRA PT. The MedDRA Lowest Level Term (LLT) will be used in defining which events are treatment-emergent. The maximum severity for each LLT during the baseline period up to first dose of the study medication will be used as baseline. If an event with missing severity is preexisting during the baseline period, and persists during the treatment period, then the baseline severity will be considered mild for determining treatment-emergence (that is, the event is treatment-emergent if the severity is coded moderate or severe postbaseline and not treatment-emergent if the severity is coded mild postbaseline). If an event occurring postbaseline has a missing severity rating, then the event is considered treatment-emergent unless the baseline rating is severe, in which case the event is not treatment-emergent. The day and time for events where onset is on the day of the first dose of study treatment will both be used to distinguish between pretreatment and posttreatment to derive treatment-emergence. Should there be insufficient data for AE start date to make this comparison (for example, the AE start year is the same as the treatment start year, but the AE start month and day are missing), the AE will be considered treatment-emergent.

In general, summaries will include the number of patients in the safety population (N), frequency of patients reporting the event (n), and relative frequency (that is, percentage; n/N*100). For any events that are gender-specific based on the displayed PT, the denominator used to compute the percentage will only include patients from the appropriate gender.

In an overview table, the number and percentage of patients in the safety population who experienced death, an SAE, any TEAE, discontinuation from the study due to an AE, permanent discontinuation from study drug due to an AE, or a severe TEAE will be summarized by treatment group.

The number and percentage of patients with TEAEs will be summarized by treatment group in 2 formats by MedDRA PT:

- nested within SOC with decreasing frequency in SOC, and events ordered within each SOC by decreasing frequency in the baricitinib 4-mg group;
- with events ordered by decreasing frequency in the baricitinib 4-mg group.

6.15.2.1. Common Adverse Events

Common TEAEs are defined as TEAEs that occurred in ≥2% (before rounding) of patients in any treatment group including placebo. The number and percentage of patients with common TEAEs will be summarized by treatment using MedDRA PT ordered by decreasing frequency in the baricitinib 4-mg group.

The number and percentage of patients with TEAEs will be summarized by maximum severity by treatment using MedDRA PT ordered by decreasing frequency in the baricitinib 4-mg group for the common TEAEs. For each patient and TEAE, the maximum severity for the MedDRA level being displayed is the maximum postbaseline severity observed from all associated LLTs mapping to that MedDRA PT.

6.15.2.2. Serious Adverse Event Analyses

Consistent with the International Conference on Harmonisation (ICH) E2A guideline (1994) and 21 Code of Federal Regulations (CFR) 312.32 (a) (2010), a SAE is any AE that results in any one of the following outcomes:

- death
- initial or prolonged inpatient hospitalization
- a life-threating experience (that is, immediate risk of dying)
- persistent or significant disability/incapacity
- congenital anomaly/birth defect

The number and percentage of patients who experienced any SAE will be summarized by treatment using MedDRA PT nested within SOC. Events will be ordered by decreasing frequency in the baricitinib 4-mg group within decreasing frequency in SOC. The SAEs will also be summarized by treatment using MedDRA PT without SOC.

An individual listing of all SAEs will be provided. A listing of deaths, if any, regardless of when they occurred during the study, will also be provided.

6.15.2.3. Other Significant Adverse Events

Other significant AEs to be summarized will provide the number and percentage of patients who:

- permanently discontinued study drug because of an AE or death

- temporarily interrupted study drug because of AE

by treatment using MedDRA PT nested within SOC. Events will be ordered by decreasing frequency in the baricitinib 4-mg group within decreasing frequency in SOC.

A summary of temporary interruptions of study drug will also be provided, showing the number of patients who experienced at least one temporary interruption and the number of temporary interruptions per patient with an interruption. Further, the duration of each temporary interruption (in days), the cumulative duration of dose interruption (in days) using basic descriptive statistics and the reason for dose interruption will be provided.

A listing of all AEs leading to permanent discontinuation from the study drug or from the study will be provided. A listing of all temporary study drug interruptions, including interruptions for reasons other than AEs, will be provided.

6.15.2.4. Criteria for Notable Patients

Patient narratives will be provided for all patients who experience certain "notable" events prior to data cutoff for the submission. See compound level safety standards for list of criteria.

6.15.3. Clinical Laboratory Evaluation

For the categorical laboratory analyses (shift and treatment emergent), the analysis period is defined as the treatment period plus up to 30 days off-drug follow-up time. The analysis period for the continuous laboratory analyses (e.g., change from baseline by time point) is defined as the treatment period excluding off-drug follow-up time.

All laboratory tests will be presented using the International Système (SI) and US conventional (CN) units. The performing central laboratory reference ranges will be used to define the low and high limits. Results pertaining to the 4 key hepatic laboratory assessments (alanine aminotransferase [ALT], aspartate aminotransferase [AST], total bilirubin, and alkaline phosphatase [ALP]) will be included as a separate analysis to address the risk of liver injury as a special safety topic (see Section 6.15.5.1).

There is one special circumstance for laboratory values to be derived based on regularly scheduled, protocol-specified analytes. The low-density lipoprotein/high-density lipoprotein (LDL/HDL) ratio will be derived as the ratio of LDL cholesterol to HDL cholesterol. There are no central lab reference ranges for the LDL/HDL ratio.

The following will be conducted for the laboratory analytes collected quantitatively:

- Box plots: Values at each visit (starting from randomization) and change from last baseline to each visit and to last postbaseline measure will be displayed in box plots for patients who have both a baseline and at least 1 postbaseline visit. The last non-missing observation in the treatment period will be used as the last observation. Individual measurements outside of reference limits will also be displayed using distinct symbols overlaying the box plot. Original-scale data will be used for the display but for some analytes (for example, immunoglobulins) a logarithmic scale may be used to aid in viewing the measures of central tendency and dispersion. Unplanned measurements will

be excluded. Descriptive summary statistics will be included below the box plot along with p-values resulting from between treatment comparison in change from last baseline to last observation. An ANCOVA model with explanatory term for treatment and the baseline value as a covariate will be used. These box plots will be used to evaluate trends over time and to assess a potential impact of outliers on central tendency summaries.

- Treatment-emergent high/low analyses: The number and percentage of patients with treatment-emergent high and low laboratory results at any time will be summarized by treatment group. Planned and unplanned measurements will be included. A treatment-emergent **high** result is defined as a change from a value less than or equal to the high limit at all baseline visits to a value greater than the high limit at any time during the treatment period. A treatment-emergent **low** result is defined as a change from a value greater than or equal to the low limit at all baseline visits to a value less than the low limit at any time during the treatment period. The Fisher's exact test will be used for the treatment comparisons. Number at risk (NAR) for the treatment-emergent high result is defined as the number of patients with a value less than or equal to the high limit at all baseline visits. NAR for the treatment-emergent low result is defined as the number of patients with a value greater than or equal to the low limit at all baseline visits for the treatment-emergent low result.

A listing of abnormal findings will be provided for laboratory analyte measurements, including qualitative measures. The listing will include but not limited to patient ID, treatment group, laboratory collection date, analyte name, and analyte finding. If needed by the safety physician/scientist, for analytes measured qualitatively, the number and percentage of patients with treatment-emergent abnormal laboratory results at any time will be summarized by treatment. Planned and unplanned measurements will be included. A treatment-emergent abnormal result is defined as a change from normal at all baseline visits to abnormal at any time postbaseline.

Note that additional analyses of certain laboratory analytes will be discussed within sub-sections of Section 6.15.5 pertaining to Special Safety topics (Section 6.15.5.1 for hepatic analytes, Section 6.15.5.2 for analytes related to hematological changes, Section 6.15.5.3 for analytes related to lipids, Section 6.15.5.4 for analytes related to renal function, and Section 6.15.5.5 for CPK).

6.15.4. Vital Signs and Other Physical Findings

For the treatment-emergent categorical analyses (shift and treatment emergent), the analysis period is defined as the treatment period plus up to 30 days off-drug follow-up time. The analysis period for the continuous analyses (e.g., change from baseline by time point) is defined as the treatment period excluding off-drug follow-up time.

Vital signs and physical characteristics include systolic blood pressure (SBP), diastolic blood pressure (DBP), pulse, weight, and BMI. Original-scale data will be analyzed. When these parameters are analyzed as continuous numerical variables, unplanned measurements will be

excluded. When these parameters are analyzed as categorical outcomes and/or treatment-emergent abnormalities, planned and unplanned measurements will be included.

The planned analyses described for the laboratory analytes in Section 6.15.3 will be used to analyze the vital signs and physical characteristics.

Table JAIY.6.8 defines the low and high baseline values as well as the criteria used to define treatment-emergence based on post-baseline values. The blood pressure and pulse rate criteria are consistent with the document *Selected Reference Limits for Pulse/Heart Rate, Arterial Blood Pressure (Including Orthostasis), and Electrocardiogram Numerical Parameters for Use in Analyses of Phase 2-4 Clinical Trials Version 1.3* approved on April 29, 2015 as recommended by the Lilly Cardiovascular Safety Advisory Committee (CVSAC).

Table JAIY.6.8. Categorical Criteria for Abnormal Treatment-Emergent Blood Pressure and Pulse Measurement, and Categorical Criteria for Weight Changes for Adults

Parameter (Units of Measure)	Low	High
Systolic Blood Pressure (mm Hg)	≤90 (low limit) and decrease from lowest value during baseline ≥20 if >90 at each baseline visit	≥140 (high limit) and increase from highest value during baseline ≥20 if <140 at each baseline visit
Diastolic Blood Pressure (mm Hg)	≤50 (low limit) and decrease from lowest value during baseline ≥10 if >50 at each baseline visit	≥90 (high limit) and increase from highest value during baseline ≥10 if <90 at each baseline visit
Pulse (beats per minute)	<50 (low limit) and decrease from lowest value during baseline ≥15 if ≥50 at each baseline visit	>100 (high limit) and increase from highest value during baseline ≥15 if ≤100 at each baseline visit
Weight (kilograms)	(Loss) decrease ≥7% from lowest value during baseline	(Gain) increase ≥7% from highest value during baseline

6.15.5. Special Safety Topics, including Adverse Events of Special Interest

In addition to general safety parameters, safety information on specific topics of special interest will also be presented. Additional special safety topics may be added as warranted. The topics outlined in this section include the protocol-specified AESI.

In general, for topics regarding safety in special groups and circumstances, patient profiles and/or patient listings, where applicable, will be provided when needed to allow medical review of the time course of cases/events, related parameters, patient demographics, study drug treatment and meaningful concomitant medication use. In addition to the safety topics for which provision or review of patient data is specified, these will be provided when summary data are insufficient to permit adequate understanding of the safety topic.

6.15.5.1. Abnormal Hepatic Tests

Analyses for abnormal hepatic tests will involve 4 laboratory analytes: ALT, AST, total bilirubin, and ALP. In addition to the analyses described in Section 6.15.3, this section describes specific analyses for this topic.

First, the number and percentage of patients with the following abnormal elevations in hepatic laboratory tests at any time will be summarized between treatment groups:

- The percentages of patients with an ALT measurement ≥3×, 5×, and 10× the central laboratory upper limit of normal (ULN) during the treatment period will be summarized for all patients with a postbaseline value and for subsets based on various levels of baseline.
    - The analysis of 3× ULN will contain 4 subsets: patients whose non-missing maximum baseline value is ≤1× ULN, patients whose maximum baseline is >1× ULN but <3× ULN, patients whose maximum baseline value is ≥3× ULN, and patients whose baseline values are missing.
    - The analysis of 5× ULN will contain 5 subsets: patients whose non-missing maximum baseline value is ≤1× ULN, patients whose maximum baseline is >1× ULN but <3× ULN, patients whose maximum baseline is ≥3× ULN but <5× ULN, patients whose maximum baseline value is ≥5× ULN, and patients whose baseline values are missing.
    - The analysis of 10× ULN will contain 6 subsets: patients whose non-missing maximum baseline value is ≤1× ULN, patients whose maximum baseline is >1× ULN but <3× ULN, patients whose maximum baseline is ≥3× ULN but <5× ULN, patients whose maximum baseline is ≥5× ULN but <10× ULN, patients whose maximum baseline value is ≥10× ULN, and patients whose baseline values are missing.
- The percentages of patients with an AST measurement ≥3×, 5×, and 10× the central laboratory ULN during the treatment period will be summarized for all patients with a postbaseline value and for subsets based on various levels of baseline. Analyses will be constructed as described above for ALT.
- The percentages of patients with a total bilirubin measurement ≥2× the central laboratory ULN during the treatment period will be summarized for all patients with a postbaseline value and subset into 4 subsets: patients whose non-missing maximum baseline value is ≤1× ULN, patients whose maximum baseline is >1× ULN but <2× ULN, patients whose maximum baseline value is ≥2× ULN, and patients whose baseline values are missing.
- The percentages of patients with an ALP measurement ≥1.5× the central laboratory ULN during the treatment period will be summarized for all patients with a postbaseline value and subset into 4 subsets: patients whose non-missing maximum baseline value is ≤1× ULN, patients whose maximum baseline is >1× ULN but <1.5× ULN, patients whose maximum baseline value is ≥1.5× ULN, and patients whose baseline values are missing.

Information collected from additional hepatic safety data collection forms will be provided in patient profiles.

Second, to further evaluate potential hepatotoxicity, an Evaluation of Drug-Induced Serious Hepatotoxicity (eDISH) plot using maximum postbaseline ALT divided by ULN vs. maximum postbaseline total bilirubin divided by ULN will be created that includes all patients from the safety populations for the studies included in the submission (any phase, any medication). Each subject with at least 1 postbaseline ALT and total bilirubin contributes 1 point to the plot. The measurements do not need to be taken at the same blood draw. Symbols may be used to indicate randomized treatment.

When criteria are met for hepatic evaluation and completion of the hepatic safety CRF, investigators are required to answer a list of questions (see Compound level safety standards). A listing of the collected information will be generated together with a graphical patient profile. This includes demographics, disposition, and a display of study drug exposure, AEs, medications, and the liver-related measurements over time will be provided for these patients and any additional patients meeting ALT or AST measurement greater than or equal to 5× ULN (on a single measurement) or ALP measurement greater than or equal to 2× ULN (on a single measurement).

6.15.5.2. Hematologic Changes

Hematologic changes will be defined based on clinical laboratory assessments. Common Terminology Criteria for Adverse Events (CTCAEs) will be applied for selected laboratory tests and are described in the compound level safety standards. These CTCAE grading schemes are consistent with both Version 3.0 and Version 4.03 of the CTCAE guidelines (CTCAE 2003, 2010).

Treatment-emergent laboratory abnormalities occurring at any time during the treatment period and shift tables of baseline to maximum grade during the treatment period will be tabulated. Planned and unplanned measurements will be included. Treatment-emergence will be characterized using the following 5 criteria (as appropriate to the grading scheme):

- any increase in postbaseline CTCAE grade from worst baseline grade
- increase to Grade 1 or above at worst postbaseline
- increase to Grade 2 or above at worst postbaseline
- increase to Grade 3 or above at worst postbaseline
- increase to Grade 4 at worst postbaseline.

Shift tables will show the number and percentage of patients based on baseline to maximum during the treatment period, with baseline depicted by the most extreme grade during the baseline period. With each shift table, a shift table summary displaying the number and percentage of patients with maximum postbaseline results will be presented by treatment group for each treatment period within the following categories:

- decreased: postbaseline category < baseline category
- increased: postbaseline category > baseline category

- same: postbaseline category = baseline category

A laboratory-based treatment-emergent outcome related to increased platelet count will be summarized in similar fashion. Treatment-emergent thrombocytosis as a laboratory-based abnormality will be defined as an increase in platelet count from a maximum baseline value ≤600 billion/L to any postbaseline value >600 billion/L (Lengfelder et al. 1998). Planned and unplanned measurements will be included.

A listing of patients with treatment-emergent thrombocytosis may be provided for safety review.

6.15.5.3. Lipids Effects

Lipid effects will be assessed through analysis of elevated total cholesterol, elevated LDL cholesterol, decreased HDL cholesterol, and elevated triglycerides as described in Section 6.15.3 and with TEAEs potentially related to hyperlipidemia.

Categorical analyses will be performed using National Cholesterol Education Program (NCEP) Adult Treatment Panel (ATP) III guidelines (2002) as shown in the compound level safety standards. The grade-like categories shown in this table are ordered from traditionally most desirable to least desirable for the purposes of these analyses.

Shift tables will show the number and percentage of patients based on baseline to the least desirable category during the treatment period, with baseline depicted by the least desirable category during the baseline period. With each shift table, a shift table summary displaying the number and percentage of patients with the least desirable postbaseline results will be presented by treatment group for each treatment period within the following categories:

- decreased: postbaseline category more desirable than baseline category,
- increased: postbaseline category less desirable than baseline category,
- same: postbaseline category = baseline category.

Treatment-emergent laboratory abnormalities related to elevated total cholesterol, elevated triglycerides, elevated LDL cholesterol, and decreased and increased HDL cholesterol occurring at any time during the treatment period will be tabulated using the NCEP categories shown in the compound level safety standards.

Treatment-emergent elevated total cholesterol will be characterized as follows:

- increase to categories ‘Borderline high’ or ‘High’
- increase to category ‘High.’

Treatment-emergent elevated triglycerides will be characterized as

- increase to categories ‘Borderline high,’ ‘High,’ or ‘Very high’
- increase to categories ‘High’ or ‘Very high’
- increase to category ‘Very high.’

Treatment-emergent elevated LDL cholesterol will be characterized as

- increase to categories ‘Borderline high,’ ‘High,’ or ‘Very high’

- increase to categories ‘High’ or ‘Very high’
- increase to ‘Very high’

Treatment-emergent abnormal HDL cholesterol will be characterized as

- decreased HDL
  - decrease to categories ‘Normal’ or ‘Low’
  - decrease to category ‘Low’
- increased HDL
  - increase to categories ‘Normal’ or ‘High’
  - increase to category ‘High’

The percentages of patients with treatment-emergent potential hyperlipidemia will be summarize by treatment, ordered by decreasing frequency in the baricitinib 4-mg group using a predefined MedDRA list of PTs that is a subset of the narrow scope PTs in the MedDRA SMQ ‘Dyslipidemia’ (code 200000026) [see Compound level safety standards].

6.15.5.4. Renal Function Effects

Effects on renal function will be assessed through analysis of elevated creatinine.

CTCAEs will be applied for laboratory tests related to renal effects as shown in the compound level safety standards. This CTCAE grading scheme is consistent with both Version 3.0 and Version 4.03 of the CTCAE guidelines. Shift tables will show the number and percentage of patients based on baseline to maximum during the treatment period, with baseline depicted by highest grade during the baseline period. Treatment-emergent laboratory abnormalities related to elevated creatinine occurring at any time during the analysis period will be tabulated. Refer to the Compound level safety standards for details.

6.15.5.5. Elevations in Creatine Phosphokinase (CPK)

Elevations in creatine phosphokinase (CPK) will be addressed using CTCAE criteria as described in the compound level safety standards. This CTCAE grading scheme is consistent with both Version 3.0 and Version 4.03 of the CTCAE guidelines. Analyses will be the same as the CTCAE analyses specified for laboratory tests related to renal function events in Section 6.15.5.2.

A listing of elevated CPK (CTCAE grade of 3 or above) will be provided for medical safety review.

Treatment-emergent adverse events potentially related to muscle symptoms may be analyzed, based on reported AEs. The Muscle Symptoms special search category is a pre-defined MedDRA search criteria list that contains the narrow scope terms from the Rhabdomyolysis / myopathy SMQ (code 20000002) plus selected terms from the Musculoskeletal SOC. These terms are shown in the compound level safety standards.

6.15.5.6. Infections

Infections will be defined using all the PTs from the Infections and Infestations SOC as defined in MedDRA. Serious infection will be defined as all the infections that meet the SAE criteria.

The number and percentage of patients with TEAEs of infections, serious infections, and infections resulting in permanent study drug discontinuation will be summarized by treatment group using MedDRA PTs. The proportion of patients developing skin infections requiring antibiotic treatment by Week 16 will also be summarized on the overview of infections table.

The number and percentage of patients with TEAEs of infections by maximum severity will be summarized by treatment group using MedDRA PTs.

Treatment-emergent infections will be reviewed in context of other clinical and laboratory parameters via a listing (details see Compound level safety standards).

The TEAE infections will be further analyzed in terms of potential opportunistic infection, herpes zoster and herpes simplex. Summary of HBV DNA monitoring results and association between infection and neutropenia/lymphopenia will also be provided in the context of infections.

Opportunistic infection

To identify potential opportunistic infections (POIs), the following approach will be used:

- identifying the POIs using a list of MedDRA PTs (refer to the compound level safety standards).

Potential opportunistic infections identified through these search approaches may be combined in one list for medical assessment and final classification of whether the case met the modified Winthrop definitions for opportunistic infections (OIs).

A final listing for OIs will be provided for the CSR and to assist the composition of patient narratives.

Herpes zoster

Cases of herpes zoster will be further classified as follows:

- localized or non-multidermatomal-involvement of the primary and/or adjacent dermatomes only
    - complicated – documented ocular (cornea or deeper structure; for example, iritis, keratitis, retinitis, etc.) or motor nerve involvement (e.g., palsy; post herpetic neuralgia [PHN] does not meet criteria for motor nerve involvement).
    - uncomplicated-localized or non-multidermatomal cases that are not complicated
- multidermatomal-involvement beyond primary and adjacent dermatomes (that is, >3 contiguous dermatomes) or involvement of two or more non-contiguous dermatomes
    - complicated-documented ocular (cornea or deeper structure; for example iritis, keratitis, retinitis, etc.) or motor nerve involvement
    - uncomplicated-multidermatomal cases
- disseminated-systemic infection, visceral or widespread cutaneous (e.g., ≥5 dermatomes or 3 to 4 dermatomes including at least 1 non-contiguous).

- Recurrent - >1 infection occurring in an individual patient during the course of participation in the baricitinib clinical program.

All herpes zoster will undergo medical review to determine the classification as described above.

A summary of herpes zoster table will be provided. The summary table will also include event maximum severity, seriousness, whether resulting in temporary study drug interruption, whether resulting in study drug discontinuation, whether treated with antiviral medication, and event outcome. The incidence rate adjusted for observation time will also be provided (as defined in Section 6.15.2). Of note, in the context of herpes zoster, antiviral medication treatment is defined as that the medication was initiated at the event start date, or within 30 days before or after the event start date. The antiviral medication for herpes zoster includes but not limited to Aciclovir, Brivudine, Cidofovir, Famciclovir, Foscarnet, Ganciclovir, Penciclovir, Valaciclovir, Valganciclovir, Vidarabine (best presented by J05AB, J05AC, J05AE, and J05AH Anatomical Therapeutic Classification codes). Medical representatives will review the concomitant medication list prior to database lock and make adjustment of the above list if necessary.

If a patient has more than 1 event of herpes zoster, the event with the maximum severity will be used in these summary tables. If more than 1 event of herpes zoster occurs with the same severity, the event with the longest duration will be used in the summary table.

Herpes simplex

A summary analysis of herpes simplex will be provided. Herpes simplex will be defined based on MedDRA PT as listed in the compound level safety standards (both narrow and broad terms in the herpes simplex section). The list needs to be reviewed by medical prior to data locks (final and interim). The summary table will include event maximum severity, seriousness, whether resulting in temporary study drug interruption, whether resulting in study drug discontinuation, and whether treated with antiviral medication.

If a patient has more than 1 event of herpes simplex, the event with the maximum severity will be used in these summary tables. If more than 1 event of herpes simplex occurs with the same severity, the event with the longest duration will be used in the summary table.

Skin Infections

A summary analysis of skin infections may be provided. Skin infections may be defined based on MedDRA preferred term (see the Compound level safety standards).

HBV DNA

A listing of patients with detectable hepatitis B virus (HBV) deoxyribonucleic acid (DNA) post baseline will be provided.

HBV DNA status post baseline (not detectable, detectable but not quantifiable [i.e., < lower limit of detection (LLOD)], quantifiable [i.e., ≥LLOD]) will be summarized by treatment group stratified by baseline HBV serology status, specifically:

- HBsAb+ / HBcAb+

- HBsAb- / HBcAb+

Association between infection and neutropenia/lymphopenia

Depending on the number of cases with CTCAE Grade 2 or greater, a summary table will be provided for treatment-emergent infections that were preceded or accompanied by neutropenia. For this analysis, neutropenia is defined as (1) CTCAE Grade 2 or greater, (2) Grade 3 or greater. Infection events with onset date ≤14 days before or after the Grade 2 or greater neutrophil count collection date will be considered as infections preceded or accompanied by neutropenia.

Similar analyses as above will be conducted to evaluate the association between infection and lymphopenia.

6.15.5.7. Major Adverse Cardiovascular Events (MACE) and Other Cardiovascular Events

Potential major adverse cardiovascular events (MACE) and other cardiovascular events requiring adjudication will be analyzed.

Categories and subcategories analyzed will include, but are not limited to, the following:

- MACE
  - Cardiovascular death,
  - Myocardial infarction (MI),
  - Stroke,
- Other cardiovascular events
  - Transient ischemic attack,
  - Hospitalization for unstable angina,
  - Hospitalization for heart failure,
  - Serious arrhythmia,
  - Resuscitated sudden death,
  - Cardiogenic shock,
  - Coronary interventions (such as coronary artery bypass surgery or percutaneous coronary intervention),
- Non-cardiovascular death,
- All-cause death.

In general, events requiring adjudication are documented by investigative sites using an endpoint reporting CRF. This CRF is then sent to the adjudication center for external adjudication which uses an adjudication reporting CRF to document the final assessment of the event as a MACE, as some other cardiovascular event, or as no event (according to the Clinical Endpoint Committee Charter). In some cases, however, the investigator may not have deemed that an event had met the endpoint criteria but the event was still sent for adjudication as a potential MACE, other cardiovascular event, or no event. These events are included in the adjudication process to ensure adequate sensitivity. In these instances, the adjudication reporting CRF will not have a matching endpoint reporting CRF from the investigator. Events generated from these

circumstances will be considered as events sent for adjudication in the absence of an investigator's endpoint reporting form.

The number and percentage of patients with MACE, other cardiovascular events, non-cardiovascular death, and all-cause death, as positively adjudicated, will be summarized by treatment group based on the categories and subcategories above.

A listing of the events sent for adjudication will be provided to include data concerning the MedDRA PT related to the event, the seriousness of the event, and the event outcome, along with the adjudicated result.

6.15.5.8. Venous and Pulmonary Artery Thromboembolic (VTE) Events

Events identified as representative of venous thromboembolic events (VTE) disease will be classified as Deep Vein Thrombosis (DVT), pulmonary embolism (PE), or other peripheral venous thrombosis and will be analyzed. The following definitions apply:

- DVT: Clinical diagnosis of a thrombosis in a deep vein above the knee that must be confirmed by objective evidence of either: a filling defect of deep veins of the leg on venography or a non-compressible venous segment on ultrasound or confirmation by other imaging modality (e.g., Computed tomography [CT], Magnetic Resonance Imaging [MRI]).
- PE: Clinical diagnosis of pulmonary embolus that must be confirmed by objective evidence of either: a filling defect of pulmonary arteries by either pulmonary angiography or CT angiography or by a high probability Ventilation Perfusion (VQ) scan.
- Other Peripheral Venous Thrombosis: Clinical diagnosis of a venous thrombosis not specified by either DVT or PE above. Other peripheral venous thrombosis disease must be confirmed by objective evidence by imaging including venography, ultrasound, CT scan, or MRI. Examples of these would include non-superficial below knee thrombosis, portal vein, subclavian vein, or mesenteric vein. Superficial thrombophlebitis alone is not considered a VTE event.

In general, events requiring adjudication are documented by investigative sites using an endpoint reporting CRF. Refer to Section 6.15.5.7 for more details as the process is the same as that of MACE.

The number and percentage of patients with a VTE, DVT/PE, DVT, PE, and other peripheral venous thrombosis, as positively adjudicated, will be summarized by treatment group. Note that the below knee thrombosis captured in the "other peripheral venous thrombosis" category will be summarized within DVT.

A listing of the VTE events sent for external adjudication will be provided to include data concerning the MedDRA PT related to the event, the seriousness of the event, and the event outcome, along with the adjudicated result.

6.15.5.9. Arterial Thromboembolic (ATE) Events

Refer to the Compound level safety standards.

6.15.5.10. Malignancies

Malignancies will be identified using terms from the malignant tumors SMQ (SMQ 20000194). Malignancies excluding non-melanoma skin cancers (NMSC) and NMSC will be reported separately.

All the cases identified by malignant tumors SMQ will be assessed through medical review to determine confirmed NMSC cases.

First, a listing including all the malignancy cases will be prepared before database lock along with the *planned* NMSC flag according to the current MedDRA version PTs (the list will be updated depending on the MedDRA version used for analysis):

- Squamous cell carcinoma of skin (10041834)
- Bowen's disease (10006059)
- Basal cell carcinoma (10004146)
- Basosquamous carcinoma (10004178)
- Basosquamous carcinoma of skin (10004179)
- Squamous cell carcinoma (10041823)
- Skin squamous cell carcinoma metastatic (10077314)
- Skin cancer (10040808)
- Carcinoma in situ of skin (10007390)
- Keratoacanthoma (10023347)
- Vulvar squamous cell hyperplasia (10079905)
- Skin squamous cell carcinoma recurrent (10081136)

This internal review is to occur prior to database lock. The case review and subsequent summary analyses will include all the cases reported in the study database or by LSS report, disregarding the length of gap between the last treatment dose date and the event date. The NMSC flag will be confirmed during the internal review process.

The number and percentage of patients with TEAEs associated malignancies excluding NMSC and NMSC will be summarized by treatment group.

6.15.5.11. Allergic Reactions/Hypersensitivities

A search will be performed using the MedDRA version 21.1 SMQs to search for relevant events, using the following queries:

- Anaphylactic reaction SMQ (20000021)
- Hypersensitivity SMQ (20000214)
- Angioedema SMQ (20000024)

Assessment of the Anaphylactic reaction SMQ includes an algorithmic query. The algorithmic approach comprises one or more events associated with an individual administration of study drug, where the events include:

- A narrow term from the SMQ (Category A of the SMQ);
- Multiple terms from the SMQ, comprising terms from at least two of the following categories from the SMQ:
    - Category B - (Upper Airway/Respiratory signs and symptoms)
    - Category C - (Angioedema/Urticaria/Pruritus/Flush signs and symptoms)
    - Category D - (Cardiovascular/Hypotension signs and symptoms).

Refer to the Compound safety level standards for details.

6.15.5.12. Gastrointestinal Perforations

Treatment-emergent adverse events related to potential gastrointestinal (GI) perforations will be analyzed using reported AEs. Identification of these events will be based on review of the PTs of the MedDRA SMQ 20000107, GI perforations (note that this SMQ holds only narrow terms and has no broad terms). Potential GI perforations identified by the above SMQ search will be provided as a listing for internal review. Each case will be assessed to determine whether it is GI perforation. A summary table based on medical review may be provided and treatment comparisons will be made using Fisher's exact test.

6.15.5.13. Columbia Suicide Severity Rating Scale

Suicidal ideation, suicidal behavior, and self-injurious behavior without suicidal intent, based on the C-SSRS, will be listed by patient and visit. Only patients that show suicidal ideation/behavior or self-injurious behavior without suicidal intent during treatment will be displayed along with all their ideation and behavior, even if not positive (i.e., if a patient's answers are all 'no' for the C-SSRS, then that patient will not be displayed). A summary of the C-SSRS categories during treatment and a shift summary in the C-SSRS categories from baseline during treatment will be provided. Refer to the Compound safety level standards for details.

6.15.5.14. Self-Harm Supplement Form and Self-Harm Follow-up Form

The Self-Harm Supplement Form is a single question to enter the number of suicidal behavior events, possible suicide behaviors, or nonsuicidal self-injurious behaviors. If the number of behavioral events is greater than zero, it will lead to the completion of the Self-Harm Follow-Up Form. The Self-Harm Follow-Up Form is a series of questions that provides a more detailed description of the behavior cases. A listing of the responses give on the Self-Harm Follow-Up Form will be provided.

6.16. Subgroup Analyses

Subgroup analyses comparing each dose of baricitinib to placebo will be performed on the ITT population at Week 16 using the primary censoring rule for the following:

- Proportion of patients achieving IGA 0 or 1
- Proportion of patients achieving EASI75 Response Rate
- Proportion of patients achieving Itch NRS 4-point improvement

The following subgroups, categorized into disease-related characteristics and demographic characteristics, will be evaluated:

- Patient Demographic and Characteristics Subgroups:
  - Gender (male, female)
  - Age group (<65, ≥65 years old)
  - Age group (<65, ≥65 to <75, ≥75 to <85, ≥85 years old)
  - Baseline weight: (<60 kg, ≥60 to <100 kg, ≥100 kg)
  - Baseline BMI (<25 kg/m^2, ≥25 to <30 kg/m^2, ≥30 kg/m^2)
  - Race: (American Indian or Alaska Native, Asian, Black or African American, Native Hawaiian or Other Pacific Islander, White, Multiple)
  - Baseline renal function status: impaired (eGFR <60 mL/min/1.73 m^2) or not impaired (eGFR ≥60 mL/min/1.73 m^2)
- Geographic Region Subgroups:
  - Region: (as defined in Table JAIY.5.1)
  - Specific regions (Europe, other)
  - Specific regions (East Asia[Korea, Japan and Taiwan], other)
  - Specific country (Japan, other)
  - Prior systemic therapy use (yes, no)
- Baseline Disease-Related Characteristics Subgroup
  - Baseline disease severity (IGA score): 3, 4

Descriptive statistics will be provided for each treatment and stratum of a subgroup as outlined, regardless of sample size. As all endpoints are categorical, subgroup analyses will be performed using logistic regression using Firth's correction to accommodate (potential) sparse response rates. The model will include the categorical outcome as the dependent variable and baseline value (for EASI and itch), baseline severity, treatment, subgroup, and treatment-by-subgroup interaction as explanatory variables. Missing data will be imputed using NRI (Section 6.4.1). The treatment-by-subgroup interaction comparing treatment groups will be tested at the 0.1 significance level. The p-value from the logistic regression model will be reported for the interaction test and the subgroup test, unless the model did not converge. Response counts and percentages will be summarized by treatment for each subgroup category. The difference in percentages and 95% CI of the difference in percentages using the Newcombe-Wilson without continuity correction will be reported. The corresponding p-value from the Fisher's exact test will also be produced.

In case any level of a subgroup comprises <10% of the overall sample size, only descriptive summary statistics will be provided for treatment arms, and no treatment group comparisons will be performed within these subgroup levels.

Additional subgroup analyses on efficacy may be performed as deemed appropriate and necessary.

6.17. Protocol Deviations

Protocol deviations will be tracked by the clinical team, and their importance will be assessed by key team members during protocol deviation review meetings. Out of all important protocol deviations (IPDs) identified, a subset occurring during Period 2 with the potential to affect efficacy analyses will result in exclusion from the PP population.

Potential examples of deviations include patients who receive excluded concomitant therapy, significant non-compliance with study medication (<80% of assigned doses taken, failure to take study medication and taking incorrect study medication), patients incorrectly enrolled in the study, and patients whose data are questionable due to significant site quality or compliance issues. Refer to a separate document for the important protocol deviations.

Trial Issue Management Plan includes the categories and subcategories of important protocol deviations and whether or not these deviations will result in the exclusion of patients from per protocol set.

The number and percentage of patients having IPD(s) will be summarized within category and subcategory of deviation by treatment group for Period 2 using the ITT population. Individual patient listings of IPDs will be provided. A summary of reasons patients were excluded from the PPS will be provided by treatment group.

6.18. Interim Analyses and Data Monitoring

An interim analyses may be conducted at the time when the last patient completes Visit 8 (Week 16) or ETV.

The baricitinib AD, AA and SLE Phase 3 programs Data Monitoring Committee (DMC) is an independent expert advisory group commissioned and charged with the responsibility of evaluating cumulative safety at regular intervals. As such, the primary objective of the DMC is to monitor the safety of the subjects enrolled in the baricitinib AD, AA and SLE Phase 3 programs by reviewing the available clinical data at scheduled time points, as described in this DMC Charter, as well as on an ad hoc basis, as needed. The DMC will consist of members external to Lilly. This DMC will follow the rules defined in the DMC charter, focusing on potential and identified risks for this molecule and for this class of compounds. Data Monitoring Committee membership will include, at a minimum, specialists with expertise in dermatology, statistics, cardiology, and other appropriate specialties.

The DMC will be authorized to review unblinded results of analyses by treatment group prior to database lock, including study discontinuation data, AEs including SAEs, clinical laboratory data, vital sign data, etc. The DMC may recommend continuation of the study, as designed; temporary suspension of enrollment; or the discontinuation of a particular dose regimen or the entire study. While the DMC may request to review efficacy data to investigate the benefit/risk relationship in the context of safety observations for ongoing patients in the study, no information regarding efficacy will be communicated. Moreover, the study will not be stopped for positive efficacy results nor will it be stopped for futility. Hence, no alpha is spent. Details of the DMC, including its operating characteristics, are documented in the Baricitinib Atopic

Dermatitis DMC charter and further details are given in the Interim Analysis Plan in Section 6.18.1.

Besides DMC members, a limited number of pre-identified individuals may gain access to the limited unblinded data, as specified in the unblinding plan, prior to the interim or final database lock, for preparation of regulatory documents. Information that may unblind the study during the analyses will not be reported to study sites or blinded study team until the study has been unblinded.

6.18.1. Interim Analysis Plan

The final analyses for this study are considered as the analyses to be presented to the DMC.

6.19. Planned Exploratory Analyses

The planned exploratory analyses are described in Sections 6.12 and 6.13. Additional exploratory analyses may be conducted such as exploring inadequate or super responders and their baseline characteristics and will be documented in a supplemental SAP. Health Technology Assessment (HTA) toolkit analyses, which may be produced, will also be documented in the supplemental SAP.

6.20. Annual Report Analyses

Annual report analyses, such as the Development Update Safety Report (DSUR), will be documented in a separate document.

6.21. Clinical Trial Registry Analyses

Additional analyses will be performed for the purpose of fulfilling the Clinical Trial Registry (CTR) requirements.

Analyses provided for the CTR requirements include a summary of AEs, provided as a dataset which will be converted to an XML file. Both SAEs and 'Other' AE are summarized: by treatment group, by MedDRA PT.

- An AE is considered 'Serious' whether or not it is a TEAE.
- An AE is considered in the 'Other' category if it is both a TEAE and is not serious. For each SAE and 'Other' AE, for each term and treatment group, the following are provided:
  - the number of participants at risk of an event
  - the number of participants who experienced each event term
  - the number of events experienced.
- Consistent with www.ClinicalTrials.gov requirements, 'Other' AEs that occur in fewer than 5% of patients/subjects in every treatment group may not be included if a 5% threshold is chosen (5% is the minimum threshold).
- AE reporting is consistent with other document disclosures for example, the CSR, manuscripts, and so forth.

Similar methods will be used to satisfy the European Clinical Trials Database (EudraCT) requirements.

7. Unblinding Plan

A separate JAIY Blinding / Unblinding Plan contains details of how the blind is maintained for this study.

8. References

Alosh M, Bretz F, Huque M. Advanced multiplicity adjustment methods in clinical trials. *Stat Med.* 2014;33(4):693-713.

Basra MK, Salek MS, Camilleri L, Sturkey R, Finlay AY. Determining the minimal clinically important difference and responsiveness of the Dermatology Life Quality Index (DLQI): further data. *Dermatology*. 2015;230(1):27-33.

Bretz F, Posch M, Glimm E, Klinglmueller F, Maurer W, Rohmeyer K. Graphical approaches for multiple comparison procedures using weighted Bonferroni, Simes, or parametric tests. *Biom J*. 2011;53(6):894-913.

Charman CR, Venn AJ, Williams, HC. The Patient-Oriented Eczema Measure: development and initial validation of a new tool for measuring atopic eczema severity from the patients' perspective. *Arch Dermatol.* 2004;140:1513-1519.

[CTCAE] Common Terminology Criteria for Adverse Events, Cancer Therapy Evaluation Program, Version 3.0, DCTD, NCI, NIH, DHHS, March 31, 2003. Available at: https://ctep.cancer.gov/protocoldevelopment/electronic_applications/docs/ctcaev3.pdf. Accessed 25 July 2017.

[CTCAE] Common Terminology Criteria for Adverse Events, Version 4.03, DHHS NIH NCI, NIH Publication No. 09-5410, June 14, 2010. Available at: https://www.eortc.be/services/doc/ctc/CTCAE_4.03_2010-06-14_QuickReference_5x7.pdf. Accessed 25 July 2017.

EuroQol Group. EQ-5D-5L User Guide. Version 2.1. Available at: http://www.euroqol.org/fileadmin/user_upload/Documenten/PDF/Folders_Flyers/EQ-5D-5L_UserGuide_2015.pdf. Accessed March 2018.

Gillings D, Koch G. The application of the principle of intention-to-treat to the analysis of clinical trials. *Drug Information Journal.* 1991;25(3):411-424.

Hanifin JM, Thurston M, Omoto M, Cherill R, Tofte SJ, Graeber M. The eczema area and severity index (EASI): assessment of reliability in atopic dermatitis. EASI Evaluator Group. *Exp Dermatol*. 2001;10(1):11-18.

Herdman M, Gudex C, Lloyd A, Janssen MF, Kind P, Parkin D, Bonsel G, Badia X. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). *Qual Life Res*. 2011;20(10):1727-1736.

Hongbo Y, Thomas CL, Harrison MA, Salek MS, Finlay AY. Translating the science of quality of life into practice: What do dermatology life quality index scores mean? *J Invest Dermatol*. 2005;125(4):659-664.

[ICH] International Conference on Harmonisation. Harmonised Tripartite Guideline: Clinical safety data management: definitions and standards for expedited reporting. E2A. 1994;Step 4. Available at: http://www.ich.org/products/guidelines/efficacy/article/efficacy-guidelines.html. Accessed 21 July 2017.

[ICH E9 R1] Estimands and Sensitivity Analysis in Clinical Trials http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E9/E9-R1EWG_Step2_Guideline_2017_0616.pdf

Khilji FA, Gonzalez M, Finlay AY. Clinical meaning of change in Dermatology Life Quality Index scores. *Br J Dermatol*. 2002;147(suppl 62):25-54.

Kimball AB, Naegeli AN, Edson-Heredia E, Lin CY, Gaich C, Nikai E, Yosipovitch G. Psychometric properties of the Itch Numeric Rating Scale in patients with moderate-to-severe plaque psoriasis. *Br J Dermatol.* 2016;175(1):157-162.

Kunz B, Oranje A, Labrèze, Stadler JF, Ring J, Taïeb A, Clinical Validation and Guidelines for the SCORAD Index: Consensus Report of the European Task Force Atopic Dermatitis. *Dermatology*. 1997;195(1):10-19.

Lengfelder E, Hochhaus A, Kronawitter U, Hoche D, Queisser W, Jahn-Eder M, Burkhardt R, Reiter A, Ansari H, Hehlmann R. Should a platelet limit of 600 x 10(9)/l be used as a diagnostic criterion in essential thrombocythaemia? An analysis of the natural course including early stages. *Br J Haematol*. 1998;100(1):15-23.

Naegeli AN, Flood E, Tucker J, Devlen J, Edson-Heredia E. The Worst Itch Numeric Rating Scale for patients with moderate to severe plaque psoriasis or psoriatic arthritis. *Int J Dermatol.* 2015;54(6):715-722.

[NCEP] National Cholesterol Education Program. Third Report of the NCEP Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel [ATP] III) Final Report, National Cholesterol Education Program, National Heart, Lung, and Blood Institute, National Institutes of Health, NIH Publication No. 02-5215, September 2002. *Circulation*. 2002;106(25):3143-421.

Program Safety Analysis Plan: Baricitinib (LY3009104) Version 5. Report on file, Eli Lilly and Company.

Protocol I4V-MC-JAIY(a). A Multicenter, Randomized, Double-Blind, Placebo-Controlled, Phase 3 Study to Evaluate the Efficacy and Safety of Baricitinib in Adult Patients with Moderate to Severe Atopic Dermatitis. Report on file, Eli Lilly and Company.

Reilly MC, Zbrozek AS, Dukes EM. The validity and reproducibility of a work productivity and activity impairment instrument. *Pharmacoeconomics*. 1993;4(5):353-365.

Schram ME, Spuls PI, Leeflang MMG, Lindeboom R, Bos JD, Schmitt J. EASI, (objective) SCORAD and POEM for atopic eczema: responsiveness and minimal clinically important difference. *Allergy.* 2012;67(1):99-106.

Snaith RP. The Hospital Anxiety and Depression Scale. *Health and Quality of Life Outcomes*. 2003;1:29.

Stalder J, Taieb A, Atherton D, Bieber P, Bonifazi E, Broberg A, Calza A, Coleman R, De Prost Y, Stalder J, Gelmetti C, Cuannetti A, Harper J, Kunz B, Lachapelle J, Langeland T, Lever R, Oranje A, Oueille-Roussel C, Revuz J, Ring J, Roujeau J, Saurat J, Song M, Tennstedt D, Van Neste D, Vieluf D, Poncet M. Severity scoring of atopic dermatitis: the SCORAD index. Consensus Report of the European Task Force on Atopic Dermatitis. *Dermatology*. 1993;186(1):23-31.

US National Archives and Records Administration. Code of Federal Regulations (CFR). Investigational New Drug Application (IND) safety reporting. 2010;Title 21:Section 312.32. Available at: https://www.ecfr.gov/cgi-bin/text-idx?SID=fd87d5f6fb00b95672dbe924a6d3b2f0&mc=true&node=se21.5.312_132&rgn=div8. Accessed 21 July 2017.

White D, Leach C, Sims R, Atkinson M, Cottrell D. Validation of the Hospital Anxiety and Depression Scale for use with adolescents. *Br J Psychiatry*. 1999;175(5):452-454.

Winthrop KL, Novosad SA, Baddley JW, Calabrese L, Chiller T, Polgreen P, Bartalesi F, Lipman M, Mariette X, Lortholary O, Weinblatt ME, Saag M, Smolen J. Opportunistic infections and biologic therapies in immune-mediated inflammatory diseases: consensus recommendations for infection reporting during clinical trials and postmarketing surveillance. *Ann Rheum Dis*. 2015;74(12):2107-2116.

Zigmond AS, Snaith RP. The Hospital Anxiety and Depression Scale. *Acta Psychiatr Scand*. 1983;67(6):361-370.